Conduct of a meta review of programme evaluations:

a case study of the SEARCH Program

Rumona Dickson

RN BN MHSc

Submitted: October 2011
Viva: November 2011
Confirmation: March 2012

This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Department of Educational Research

Lancaster University

Declaration

This thesis is entirely my own work and has not been offered previously for any other degree or diploma

Signature

Conduct of a meta review of programme evaluations: a case study of the SEARCH Program

Abstract

This thesis presents a retrospective case study that critically examines the evaluations that were undertaken as part of a continuing professional development (CPD) programme for health care professionals. The case is the SEARCH Program, an innovative CPD programme, which was designed to promote the implementation of evidence based practice (EBP) within the existing health care system in Alberta, Canada.

Two approaches from the 'using' branch of Alkin and Chrisite's evaluation theory tree are used in this research. The first employs a quantitative metaevaluation tool to retrospectively assess the quality of evaluations that were conducted from 2000 to 2005. The second is qualitative and explores the use of evaluations to inform programme development.

The results of the quantitative analysis demonstrate that the evaluations scored poorly. In fact all evaluations failed to meet basic pass/fail criteria in three of the four standard categories. Reasons for this are explored and include the interdependence of criteria in the metaevaluation tool, the poor or incomplete quality of the reports and the retrospective nature of the process that did not allow for additional data collection. The apparent precision offered by the metaevaluation tool is questionable, as there is a lack of explanation regarding the weighting of the various items, the quantitative formulae used, and the criteria for classifying an evaluation as a failure. The tool is also limited by its focus on evaluation process with no consideration given to the results of the programme evaluations.

The application of qualitative method was also time consuming but more fruitful. The results of the qualitative analysis demonstrate that the SEARCH Program was a complex, innovative and evolving programme functioning in a complex and changing health care system. Evaluation processes used within the programme were developmental in nature and informed substantive programme changes. The

extent of the changes extend beyond what would be expected with standard formative or summative evaluation and fit with the concepts and use of developmental evaluation as articulated by Patton.

The development of CPD programmes for health care professionals who are required to implement EBP is complex and requires collaboration between networks of professionals from institutions within health and higher education. Such programmes need to be reflective, innovative and flexible in nature due to the complex environments in which they are established and the complex outcomes that they wish to implement. This complexity and need for consistent re-evaluation of the goals of the programmes means that developmental evaluation may be an appropriate approach.

It is acknowledged that developmental evaluation is difficult and requires both expertise and commitment of those involved. It is also acknowledged that such evaluation may be able to demonstrate immediate outcomes of the CPD programme for the participants and even the faculty but is much less likely to be able to demonstrate impact on the health care system in which it is used.

Acknowledgements

I want to thank my supervisors, Professors Paul Trowler and Murray Saunders for their guidance during this PhD journey. I would also like to thank Professor Malcolm Tight and Dr. Paul Ashwin for their contributions as part of the Department of Educational Research Programme and of course Alison Sedgwick for keeping it all on track. In addition I would like to thank my Canadian academic mentors Professor Margaret Edwards and Dr. Ann Casebeer.

I want to acknowledge the invaluable support of the 'Red Robe Club'. Dr. Ian Willis, Dr. Julie Williams and Denise Cormack made up this study support group and their insights, humour and shared experiences made this a more valuable and enjoyable journey. I also want to acknowledge the staff of LRiG who experienced this journey with me and contributed in so many ways to the outcome.

I would be remiss if I did not thank all those people in the SEARCH Program that allowed me to share their experiences and their journeys.

Finally, to all the friends and family, on two continents who provided moral and logistic support – thank you and now we can have the promised celebration.

Table of contents

1 INTRODUCTION

The majority of doctoral projects are a journey and this one is no exception. It began with my introduction to evidence based practice (EBP) in the late 1980s. It moved into synthesising health research evidence and then to teaching EBP and research synthesis in institutes of higher education. There was then just the final part of the journey, how do you evaluate that teaching in terms of the students and ultimately its possible impact on the delivery of health care?

In this introductory chapter I briefly outline this journey to provide context. I go on to present the aim and research questions that guided the research project, the theoretical perspective used and provide an overview of this thesis.

1.1 The journey

Evidence based practice in health care is a concept born from work done at McMaster University in Canada in the mid and late 1980s (Sackett et al., 1997) and advocated later in the UK by others (Chalmers and Altman, 1995, Muir Gray, 1997). The concepts were not totally new and the idea that the findings from research should be used to inform clinical practice had been advocated earlier in the UK by Archie Cochrane (Cochrane Collaboration, 2010). There are three key aspects to EBP; what evidence should be used, how should the findings from multiple studies be synthesised and once a decision is made regarding best practice how can changes be implemented in health care policy and practice? Clearly there was a need for changes in the approaches used in clinical practice but there was no clear idea how to move this important policy and practice agenda forward. I was a masters student at McMaster when this movement was beginning.

In the mid 1990s I was fortunate enough to work with one of the original groups in the UK that was synthesising health research evidence to inform national health policy. I later moved to doing similar work but in the international arena in the area of infectious diseases in developing countries. This second position also brought me to teaching EBP. This began the final part of the journey that led to a desire to gain a better understanding of how to evaluate such teaching programmes, not just from the perspective of the knowledge gained by the

students but in the wider context of the impact on the delivery of health care services.

In 1998 I was introduced to the Search Program.[1] SEARCH originally stood for 'Swift, Efficient Application of Research in Community Health', however from the earliest days it was known only as the SEARCH Program. It was a programme developed in Alberta to address the implementation issues related to EBP. It was an innovative, collaborative, interdisciplinary continuing professional development (CPD) programme. Its overall purpose was to build the capacity of those working in the Alberta health care system by supporting quality decision making based on appropriate evidence (SEARCH Canada, 2007). The SEARCH Program, its evaluations and documentation have provided the data for this thesis which was designed to explore aspects of education programme evaluation. This research is not an evaluation of SEARCH but falls within the genre of research into evaluative practice (Saunders et al., 2011). The research has been guided by the following aim and research questions.

1.2 Research aim and questions

1.2.1 Research aim

To critically examine and assess the applicability, use and practices associated with evaluation within the context of programme documentation and programme evaluations related to a continuing professional development programme for health care professionals.

1.2.2 Research questions

1. What was the quality of the programme evaluations conducted during the existence of the programme when assessed using international quantitative standards for programme evaluation?
2. What role did programme evaluations play in the development and evolution of the SEARCH Program?

[1]The nature of this thesis has proved a dilemma in the use of North American versus British terminology and spelling. For the text of the document the British spelling is used. However when terms are attached to specific titles (e.g. the SEARCH Program) or used in quotes or the reference list the North American spellings are used.

3. What implications might this have for the evaluation of future continuing professional development programmes?

The aim and research questions were addressed through a retrospective case study approach that examined the extensive evaluations and other programme records of the SEARCH Program. These have been critically examined through the two different lenses of metaevaluation and developmental evaluation.

Metaevaluation is well known and comes with a set of internationally accepted standards for assessment (Joint Committee on Standards for Educational Evaluation, 1994, Yarbrough et al., 2011). These standards can be applied to completed evaluations and the quality of these evaluations can be judged through a previously designed assessment tool (Stufflebeam, 1999).

The role of evaluations on programme development is a somewhat more difficult area to examine. Historically, evaluation has been viewed in terms of formative and summative evaluation. In this dichotomy formative evaluation has played the role of examination of programmes with a purpose of informing programme changes where appropriate. More recently, a specific designation of 'developmental evaluation' has emerged, which claims to go beyond the boundaries of formative evaluation (Patton, 2011).

Developmental evaluation has yet to be researched in any depth (Gamble, 2008, Patton, 2011). Patton (Patton, 2011) says that developmental evaluation '*guides action and adaptation in innovative initiatives facing high uncertainty'*(pg36). Gamble (Gamble, 2008) outlines that developmental evaluation is most appropriately used in situations where there is high complexity and the innovations are taking place in a new or early stage of social innovation where there is likely to be significant change taking place (Gamble, 2008). It is argued in this thesis that this is the type of situation in which the SEARCH Program was conceived and implemented.

This thesis uses programme evaluations and documents from the SEARCH Program archives to explore issues related to the quality assessment of

programme evaluations and to contribute to the emerging discussion regarding the applicability of the use and concepts of developmental evaluation.

1.3 Theoretical context

Two approaches are used and are contextualised within what is a relatively new and evolving area of evaluation theory (Alkin and Christie, 2004, Christie and Alkin, 2008). The theory categorises programme evaluation approaches into three branches of a theory tree; use, methods and valuing. Both metaevaluation and developmental evaluation reside in the use branch of the evaluation theory tree. These theoretical concepts are discussed more fully later in this thesis.

1.4 Thesis structure

The following chapter provides the context of the SEARCH Program and presents in more detail. An overview of the literature follows and presents a summary of the theories and approaches used in programme evaluation and provides rationale for the choice of the two lenses used in this research. Chapter four goes on to outline the methods that were employed to examine the SEARCH Program evaluations and records through the two chosen lenses, while chapter five presents the results obtained. Chapter six presents a discussion of the findings, while the final chapter brings together the findings and provides conclusions and implications for practice.

2 SEARCH PROGRAM CONTEXT

Continuing professional development has historically been a mandatory part of the professional practice of all healthcare professionals (Murphy et al., 2006, Nursing in Practice, 2010). However, the introduction of EBP in the late 1980s led to a shift in the focus, content and delivery of such programmes. There was an identified need for these programmes to include not just the findings of current relevant research, but also to provide health professionals with the opportunities to develop the skills necessary to identify, quality appraise, synthesise and, where appropriate, incorporate the relevant research findings into both health policy and clinical practice. These topics were not historically included in established professional education programmes and had not yet been incorporated into CPD offerings (Hamer and Collinson, 2005).

A number of CPD models were in use at this time, including full and part-time delivery, credit and non-credit. However, as in other areas of education, few of these models had been formally evaluated and none had been designed to deliver the content required to meet the requirements of EBP. A leader in this field in Alberta, Canada took on the challenge of developing, delivering and evaluating an innovative CPD model to address these issues with a programme designed to meet both individual and provincial healthcare delivery needs. The SEARCH Program was a multi-disciplinary education programme that was organised in two year cohorts (e.g. SEARCH I, II, III etc.). As will be seen later, unlike the majority of CPD programmes the SEARCH Program was extensively evaluated over a period of 14 years, with evaluations conducted at session, module, cohort and programme levels.

The following sections provide the context of the programme and include an outline of the overall healthcare delivery system in Canada and then go on to describe the organisational structure and accomplishments of the SEARCH Program during its 14 year history.

2.1 Healthcare in Canada

This section is not meant to provide a detailed description of the Canadian healthcare system. However some relevant background is necessary to enable the reader to understand the development of the SEARCH Program in context.

Canadians have the benefit of a universal healthcare insurance system and Canadians (in the ten provinces and three territories) are entitled to access healthcare services. This does not mean that everyone receives the same care. As is the case in other healthcare systems, individuals are able to purchase additional private care and there are local and regional differences in the delivery of care (what has come to be known as postcode differences). The key feature of the Canadian healthcare system that was critical to the development of the SEARCH Program is that healthcare in the country is a provincial/territorial responsibility. As a result, even though there is a national health minister, all pertinent decisions related to healthcare are taken at the provincial or territorial level. This system has advantages in that local decisions can be made to meet local needs. However it also has disadvantages in that there is a lack of national strategy or consistent implementation and delivery of care.

It is within this context then that provinces across the country developed different approaches to the delivery of CPD training when faced with the need to increase capacity and systems to move forward with the issues raised in relation to the implementation of EBP. It is also within this context that a 'made in Alberta' programme evolved.

2.2 SEARCH Program

The SEARCH program was the vision of Dr. Mathew Spence, the Director of The Alberta Heritage Foundation for Medical Research (AHFMR). The foundation was established as a corporation of the Government of Alberta in 1980 and is governed by an appointed Board of Trustees. As such it is an autonomous body but adheres to the regulations of the province. Its stated objective is: '*to establish and support a balanced long-term program of medical research based in Alberta directed to the discovery of new knowledge and the application of that knowledge to improve health and the quality of health services in Alberta.'* (Alberta Heritage

Foundation for Medical Research, 2004). As such the foundation supported a broad range of research activities that included biological (laboratory based), clinical (clinical trials) and health services research.

The SEARCH Program was therefore a partnership programme that included collaboration between AHFMR, regional and provincial authorities, universities and government. When the SEARCH Program was initially conceived, the programme was modelled on an existing international programme with similar capacity-building goals. The International Clinical Epidemiology Network (INCLEN) Program was established by the Rockefeller Foundation in 1980 and, as an international collaboration designed '*to strengthen national healthcare systems and improve health practices globally by providing professionals in the field with the tools to analyze the efficacy, efficiency, and equity of interventions and preventive measures* (International Clinical Epidemiology Network, 2010).

Designed to build capacity in the healthcare system for producing and using research evidence to support healthcare planning and management decisions, the SEARCH Program provided an opportunity to develop local expertise for collaborative applied health research and evidence-based decision-making.

Program goals as stated at the time of program inception in 1996 (Birdsell and Mathias, 2001) were:

1. To have health professionals in the health authorities and agencies use current, relevant and appropriate information to assist in identifying priority health issues and in making decisions on these issues based on research results.
2. To develop a collaborative network of expertise across Alberta to initiate and carry out health research on a local, regional, or provincial basis.
3. To create a culture in which policy-responsive research is both valued and supported. (pg 7)

These goals were restated in 2001 to reflect the curriculum design used by the SEARCH Program as well as including emphasis on evaluation of the programme: (Birdsell and Mathias, 2001):

> CREATING EVIDENCE: to develop a collaborative network of expertise across Alberta to initiate and carry out health research on a local, regional or provincial basis.
>
> CHOOSING AND USING EVIDENCE: to have health professionals use current, relevant and appropriate information to assist in identifying priority health issues and making decisions based on research results.
>
> ADDING TO WHAT WE KNOW: to evaluate and further develop the SEARCH Program.
>
> CHANGING THE CONTEXT: to create a culture in which policy-responsive research is both valued and supported. (pg 7)

In a later report, Birdsell et al (2005) elaborate on each of these points but indicate that the global aims of the programme had not changed since the programme's inception.

2.3 SEARCH Program structure

The delivery of the SEARCH Program evolved over time. In general, SEARCH participants were recruited from health regions and provincial health authorities interested in the implementation of EBP; these agencies served as collaborators in the programme and sponsors of the participants. Methods of participant selection varied across the sponsors (open competition, volunteers, appointment). The sponsorship commitment of the employers included the release of the SEARCH participants from their work responsibilities for 25-80% of their work time over a two-year period. This time included attendance at residential teaching modules, carrying out one internal project (jointly determined by the sponsor, the SEARCH participants and the SEARCH faculty) and collaborating on one joint provincial project with other SEARCH participants. Twenty-five participants were recruited for each two year SEARCH cohort. Six cohorts completed the programme.

SEARCH Program delivery included a number of facets: face-to-face residential modules, inter-module activities, individual and group projects, the Desktop (integrated on-line resource centre), integrated curriculum, faculty team support, the SEARCH network and SEARCH manager support. The programme content focused on the three inter-related components of choosing, creating and using

research. Figure 1 gives an overview of the curriculum frame with details of the contents. Appendix 1 provides a sample of the curriculum themes and a programme agenda for one week of a SEARCH Program residential module.

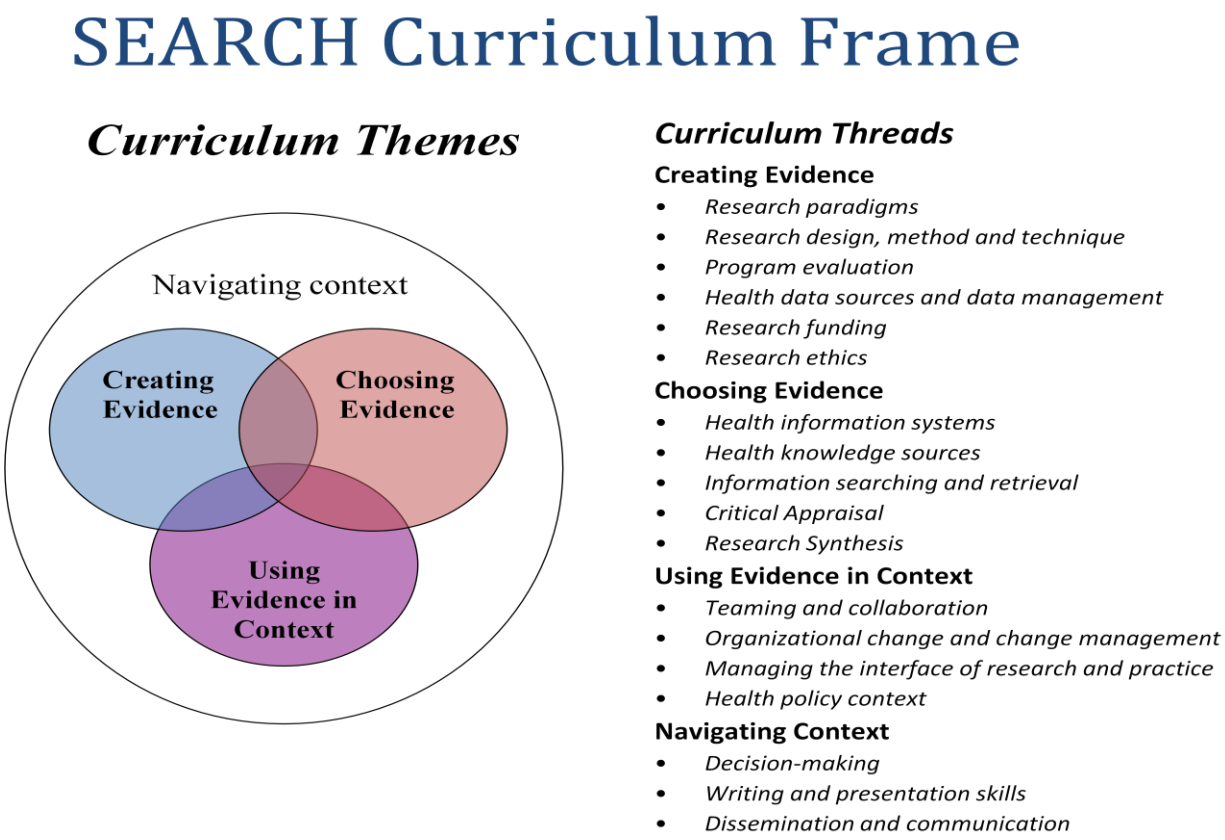


Figure 1 SEARCH Curriculum Frame

A key decision was taken early in the development of the programme to provide access to the most up-to-date computer technologies and to support students in the use of these technologies. This included the provision of laptop computers to all students and internet access to each other and to library and search facilities. If this were happening today no one would be surprised. However, in 1996, this was very innovative. The students had use of laptops and networks that were not yet available in the institutions in which they were working. A number of SEARCH participants in the first two cohorts were the only health professionals in their health region with access to the internet at work.

The SEARCH Program was founded on the basis of EBP and as such there was a strong commitment to the use of evidence and research to inform its formation,

development and impact. In addition the SEARCH participants were unique individuals, in that they were leaders in their fields, interested in the implementation of EBP and therefore were willing participants in ongoing research and evaluation of the programme and their own professional fields. Therefore extensive internal and external evaluations were undertaken throughout the programme. Methods varied and included both quantitative and qualitative approaches to data analysis. These are presented in more detail later in this thesis.

The programme included six full cohorts of students and ran from 1996 until its sudden termination by the Minister of Health in June 2009. Filed correspondence for the previous 12 to 18 months indicated that there were issues in securing funding from stakeholders and that the termination was not the result of any unsatisfactory programme evaluation but as part of the government's response to Canada's economic crisis and significant healthcare system upheaval.

2.4 Overview of SEARCH evaluations

As noted above the SEARCH Program and faculty were dedicated to the use of evaluation to inform programme development and to assess programme impacts. Formative and summative evaluations were conducted to assess all aspects of the programme including programme and curriculum design, training modules, short and long term impact on participants, teaching and learning strategies and research network development and impact. An evaluation framework was established at the inception of the programme in 1996. Although a copy of this framework was not available in the programme archives, an overview of the framework was presented in a 2003 document that outlined the evaluations conducted from 1996 until 2000 as part of that framework (Hayward, 2003).

In 2001, as the result of a SEARCH facilitated workshop, an 'evaluation blueprint' was developed that established the evaluation plan for the following 15 years. (Birdsell and Mathias, 2001). This document closely follows the recommendations set out by Saunders (2000) in his guide to evaluation planning in that it outlined the purpose, audience, principles and foci for future SEARCH evaluations. This framework clearly demonstrated a commitment for programme

evaluation to be broad and include impact on students, faculty, health organisations and the provincial health system.

In terms of purpose the blueprint document states that evaluations should be designed (Birdsell and Mathias, 2001);

> *1. To determine if SEARCH fills an unmet need in Alberta.*
>
> *2. To determine if SEARCH, as delivered, meets its program goals and contributes to the missions of the participating organisations. If not, the evaluation is designed to provide information to improve SEARCH design or implementation.*
>
> *3. To contribute to determination of whether the SEARCH concept or key attributes are transferable to other settings.*
>
> *4. To build capacity for research in practice through the design and implementation of the evaluation and research projects.* (pg 2)

The identified audiences for the evaluations were broad, and ranged from the staff in health regions, to the AHFMR board, SEARCH faculty and participants, academic institutions and contributors to possible future programmes.

The principles laid out in the blueprint are important because they guide the use, design and conduct of future evaluations (Birdsell and Mathias, 2001).

> *1. The evaluation products should be used to add to the body of knowledge of fields that inform similar programs.*
>
> *2. Evaluations will be designed, planned and implemented by an appropriate balance of knowledgeable insiders and uninvolved but experienced knowledgeable outsiders who are able to view the SEARCH program in context, critically.*
>
> *3. The evaluation activities should model best practices in evidence-based decision-making related to evaluation and program design.*
>
> *4. The evaluation activities themselves contribute to SEARCH goals, and are conducted in ways that embody the SEARCH principles.* (pg 3)

The foci for the evaluations considered programme goals, core values and beliefs, purpose and the mandate of participating partners and impact of the programme. A broad overview of the areas evaluated and methods used for evaluation is presented in Table 1.

As can be seen from the information in Table 1 the commitment to evaluation extended to all levels of the programme. Each module was evaluated through the use of talking circles, surveys and in some cases follow-up interviews and focus groups, while the most extensive evaluation was the cohort longitudinal follow-up which tracked participants over time as they completed the SEARCH Program and moved on with their respective careers.

Table 1 Evaluation focus and methods used by SEARCH Program

Sessions	**Module evaluation**	**Cohort evaluation**	**Programme Evaluation**	**Impact**
Verbal in class feedback Electronic evaluation forms	Pre-post talking circles Written evaluation form Follow-up evaluation with surveys, interviews and focus groups	Email surveys Focus groups Telephone interviews Done initially at 4, 6, 12, 18 and 24 months	On-line/email surveys Interviews Focus groups Workshop Faculty evaluation	Surveys Interviews Workshops

An indication of the commitment to evaluation is the fact that every teaching session was evaluated. The early adoption of computer technology meant that every lecture/session in every SEARCH module was evaluated through completion of a computer generated evaluation form. Anecdotal evidence from SEARCH faculty indicated that in the early days of SEARCH – that is in SEARCH I and II cohorts, these evaluations were examined in a faculty meeting at the end of each day and, where appropriate, changes were made to sessions planned for the next day or sessions planned for future modules. As technology improved, session evaluations were automatically sent by email to the presenter within 30 minutes of completion of his/her session. In today's technology this is seen as normal practice, however in the late 1990s it was seen as leading edge. In fact it is unlikely that such a comprehensive system of evaluation and feedback would be typical of programmes today even though the technology is more advanced and readily available.

As noted in the evaluation principles listed above and as will be seen later in this document, evaluations were conducted by both internal faculty members and external consultants.

In summary the SEARCH Program had broad overarching aims, involved a wide spectrum of stakeholders (participants, health regions, AHFMR) and utilised a new and evolving model of CPD programme delivery. There was, as yet, no consensus about what information the participants required to be able to implement EBP. There was, and still is, limited evidence to indicate what methods work in the process of implementing EBP, as seen in the systematic reviews conducted by the Cochrane Collaboration Effective Practice and Organisation of Care Group (Cochrane Collaboration, 2010). Given the complexities inherent in the SEARCH Program, the evaluation of it was never going to be straightforward.

3 LITERATURE OVERVIEW

This chapter provides an overview of the evolving theory of programme evaluation and a description of the methods recommended for evaluating evaluations (metaevaluation) as used to inform the conduct of this research. This overview is not the result of a systematic search or a comprehensive review of the literature and therefore does not present information on the search strategies used or criteria used for inclusion of the data presented (Dickson, 2005).

3.1 Education programme evaluation theory

Cronbach et al (1980) provide an historical perspective of the evolution of evaluation in the USA with a focus on the evaluation of new social programmes; and in their book they review 95 theses related to evaluation. They portray evaluation as an exciting and evolving field as evidenced in this quote. '*Evaluation has become the liveliest frontier of American social science. It invites-even entices-members of traditional disciplines to leave their settled fields and migrate to a land where history is being made*.'(pg 13) This accounts for the attraction for researchers from a broad range of areas – including political scientists and economists interested in social processes who could then collaborate with sociologists, anthropologists and psychologists with expertise in data collection. It is interesting to note that very few of 95 theses listed by Cronbach et al (1980) included the kind of stakeholder involvement that is included in the evaluation perspectives that are examined in depth later in this thesis.

One might question the enthusiasm of Cronbach (1980) and his associates for evaluation. It is worth keeping in mind that the group was based in Stanford California. California at the time the centre of evaluation of all kinds – most US market testing was done there because of the diversity of the population. A common jibe of the 1970s was that '*God populated California by picking up the USA by New York and shaking it so that all the loose bits ended up in California*.' Having said that, significant social research endeavours started there and the work of Cronbach et al (1980) was no exception as it broadened the approaches used and the individuals involved in evaluation.

Given this background it was therefore somewhat surprising to find that although programme evaluation discussions date back to the middle of the 20th century, the first attempt to bring this under the umbrella of evaluation theory appears to date from work done by Alkin and Christie (2004) and re-visited by them in 2008 (Christie and Alkin, 2008).

Their evaluation theory tree (Figure 1) is based on a dual foundation of systematic social inquiry and social accountability and fiscal control. The tree is formed by three main branches of evaluation; methods, valuing and use.

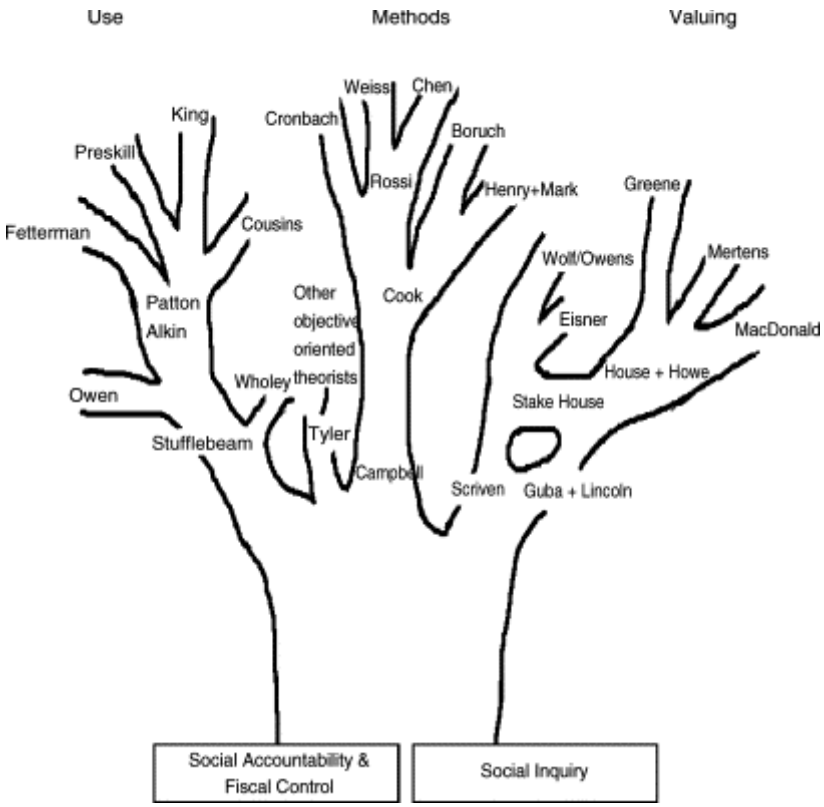


Figure 2 Christie and Alkin Evaluation Theory Tree (2008)

The first paper (Alkin and Christie, 2004) provides a limited explanation of the foundation of the theory tree. However, it provides extensive detail of each of the branches, reasons for the positioning of various theorists on each of the branches and an overview of each of the individual theories. The second paper primarily provides details of changes that have been made following the authors' reflections as well as feedback from the theorists themselves. These three branches are briefly

described here and although they are described as distinct the reflections of the various theorists point out that there is significant overlap.

Methods

The authors outline how methods of evaluation dominated the early evaluation research field and date this back to the work of Donald Campbell (Shadish and Luellen, 2004). It is based on the positivist perspective that promoted the use of experimentation and thus the definition of objectives and outcomes in evaluation.

Valuing

The values branch has as its beginning the work of Michael Scriven (Scriven, 1997, Scriven, 2005) and is focussed on valuing and making judgements about programmes and the use of techniques such as 'goal-free' evaluation. In re-visiting their evaluation tree, Christie and Alkin (Christie and Alkin, 2008) divide the valuing branch into sub-branches representing constructivist theories and post-positivist approaches. They admit that this branch of their theory tree has been the most difficult to define owing to the diversity of approaches.

Use

The final branch, and the one which informs this research, is the 'use' branch. The focus of evaluation on this branch is the need for the evaluation to be of use to programme stakeholders as they make decisions regarding continuing with or making changes to existing programmes. The views of two theorists from this branch Stufflebeam (1974, 1999) and Patton (2011) are used in the research project presented in this thesis. Each of their views is described in more detail later in this chapter.

3.2 Programme evaluation perspectives

Stufflebeam et al (2000) compiled a comprehensive overview of educational evaluation models in a book that includes contributions from leaders in the field of evaluation. They first provide an historical overview of the evolution of the role of educational evaluation from school accreditation to what they now describe as a 'maturing discipline' that is being used across a variety of sectors from education, community development, government and education (Madaus and

Stufflebeam, 2000). In the same chapter Madaus and Stufflebeam (2000) point out that programme evaluation is not a recent development. They go on to provide an extensive overview in the context of seven different periods in history beginning in 1792. The last three periods they identify as the *Age of Innocence* (1958-1972), *Age of Development* (1973-1983) and the *Age of Expansion and Integration* (1983-2000), indicating the most recent changes in thinking and approach to evaluation have taken place in these last three periods.

The first of these periods covers early attempts at educational programme evaluation carried out by people such as Ralph Tyler. The next period includes the refining of the evaluation process while the final period saw the recognition of the need to evaluate the evaluation process.

Madaus and Kellaghan (2000) present an overview of useful metaphors that have informed evaluation beginning with the '*factory model'* and continuing with '*schooling as travel'*. In the first the curriculum is seen as the means of production, while in the second the student is the raw material that is moulded by the teacher, but in both, the outcome can identified and measured with set quality criteria. However the metaphors are limited given the reality of schools and the multiplicity of expected outcomes. The travelling metaphor links more to education as a lifelong journey during which the student travels and is aided by various external resources along the way – including having the teacher as a guide and fellow traveller. Obviously the approach taken by the teacher in each of these situations is different as are the measured outcomes.

Madaus and Kellaghan (2000) go on to argue that these various perspectives on education have informed subsequent evaluation approaches. They report the result of their culling of writings from evaluation theorists past and present and group their findings under 20 evaluation definitions that range from objective/goal based, through legal, to naturalistic. They do not consider their list as comprehensive and acknowledge the limitations of the single definition used for each category.

Importantly, they also introduce and define the terms ‘merit’ and ‘worth’ with respect to evaluation (Madaus and Kellaghan, 2000).

> *Merit: The excellence of an object as assessed by its intrinsic qualities of performance*
>
> *Worth: The value of an object in relationship to a purpose.* (pg 29)

In the same volume Stufflebeam (2000a) identifies and classifies 22 different evaluation approaches. He clearly states that the decisions regarding these classifications are based on his experience and judgement but does outline the historical premises on which they are based. He divides his 22 approaches into four categories. The first includes two approaches which he calls pseudo-evaluations; they encompass what he defines as public-relations inspired and politically controlled evaluation. It is quite clear that he sees both approaches as presenting invalid or incomplete findings and they are given very little further attention. The remaining 20 approaches span the other three categories: improvement/accountability; social mission/advocacy and questions/methods.

Improvement/accountability evaluations (n=3) consider programme merit and worth; they are comprehensive, generally objective/quantitative in nature and designed both to improve programmes and provide consumers with information about and access to those programmes. Social mission/advocacy evaluations (n=4) focus on the importance of universal access to programmes. The largest category, questions/method (n=13) includes assessment of merit and worth but with a focus on comparison to set of accepted programme standards and are frequently qualitative in nature.

Although these categorisations are helpful, Stufflebeam more importantly provides an analysis that rates each of the approaches in relation to its *potential ability* to be assessed using the Joint Committee Program Evaluation Standards, which assess evaluations in terms of their utility, feasibility, propriety and accuracy (Joint Committee on Standards for Educational Evaluation, 1994, Stufflebeam, 1999).

The results of this analysis leave only nine approaches that are rated very good or good with respect to their potential to meet the standards. These approaches are presented in Table 2 with an explanation of each.

Table 2 Evaluation approaches rated good/very good

Evaluation approach	Characteristics
IMPROVEMENT ACCOUNTABILITY	
Decision/accountability	Retrospectively assesses merit and worth as well as proactively informing programme improvement
Consumer oriented	Assess merit and worth in the context of consumers' welfare
Accreditation	Meeting of pre-set standards (e.g. hospital accreditation)
SOCIAL MISSION/ADVOCACY	
Utilisation-focused	Stakeholder focused with an emphasis on how the results of the evaluation are used
Client-centred	Designed to focus on those who plan and deliver the programme
Democratic deliberative	Democratic framework to ensure democratic principles are upheld allowing for input from all stakeholders
Constructivist	Philosophical, service oriented and paradigm driven. Evaluators' role is to manipulate the evaluation to emancipate and empower the disenfranchised
QUESTIONS/METHODS	
Case study	Focused on in-depth description, analysis and synthesis of a particular programme
Outcome monitoring/value added	A special case that uses standardised testing as well as examining overall results that can be compared across centres

Summarised from Stufflebeam (2000a)

A different perspective on evaluation is taken by Chelimsky (Chelimsky, 1997), a leader in the field of evaluation research. Her view is that evaluation is driven by three, sometimes overlapping, goals. These are:

- *Evaluation for accountability (e.g. the measurement of results or efficiency)*
- *Evaluation for development (e.g. the provision of evaluative help to strengthen institutions)*
- *Evaluation for knowledge (e.g. the acquisition of a more profound understanding in some specific area or field (pg10)*

She then goes on to present these three perspectives in relation to nine dimensions of evaluation. The use of these dimensions, although not as detailed or as

quantitative as those outlined by Stufflebeam (2000a), provides a framework for the critical appraisal of a given evaluation. These perspectives are presented in Table 3. As can be seen there is significant overlap with the categories set out by Stufflebeam (2000a) above.

Table 3 Three perspectives and their positions in nine dimensions

Dimension	Accountability perspective	Knowledge perspective	Developmental Perspective
Purpose	To measure results or value for funds expended To determine costs To assess efficiency	To generate insights about public problems, policies, programs and processes To develop new methods and to critique old ones	To strengthen institutions; To build agency or organisational capability in some evaluative area
Need for use to fulfil purpose	No	No	Yes
Typical uses	Policy use Debate and negotiation Enlightenment Governmental/agency reform Public use	Enlightenment use Policy Research and replication Education Knowledge base construction	Institutional or agency use as part of the evaluative process Public and policy use
Evaluator role re client	Distant	Distant or close depending on evaluation design and methods	Close; Evaluator is a 'critical friend' or may be part of a team
Independence	Prerequisite	Critical	Little need
Advocacy	Unacceptable	Currently unacceptable, but now being debated	Often inevitable, but correctable through independent, outside review
Acceptability to clients or users	Often difficult but may be helped by negotiation	Clients may ignore or shelve findings they do not like	Easy: no threat posed
Objectivity	High	High (when advocacy is not present)	Uncertain (based on independence and control
Position under policy debate	Can be strong (depending on leadership)	Can be strong (if consolidated and dissemination channels exist)	Uncertain (based on independence and control)

Adopted from (Chelimsky, 1997) pg 21

Given the evolving nature of education programme evaluation it is not surprising that a variety of methods are used and that there is on-going discussion regarding the role of the evaluator and the evaluation. In the final two chapters of their book, Chelimsky and Shadish (1997) present views from two theorists representing extreme positions regarding programme evaluation.

Stake (1997) outlines the importance of the role of the evaluator as an advocate. He uses as his example an evaluation that he has conducted in which he was intimately involved with the group being evaluated and argues for the need to advocate and even protect the programme and the individuals being evaluated. He does this through what is now known as a form of bias in randomised controlled trials (RCT) and meta-analysis called 'selective reporting' (Dwan, 2010). That is, he selectively reports different findings to different groups (the public, the

programme administrators and the programme subjects). He argues that this is valid because there is fear that a negative evaluation will lead to the closure of an important programme that is providing valuable services not provided elsewhere and that even though there is room for improvement, the services that are provided are needed and important. Had he presented any information regarding working collaboratively with the stakeholders in this project then this might have been considered a utilisation-focused evaluation (Patton, 2008). However, no indication of collaboration is discussed.

Scriven (1997) on the other hand takes a positivist view of the role of evaluation. He argues that there are objective truths regarding any programme being evaluated and it is the role of the evaluator to report these truths. He argues for the maintenance of distance between the evaluator and those being evaluated to decrease/limit the biases that might be caused by *'personality clashes, personal attraction, and other personal feelings...'*(pg 481). He maintains that this objectivity and distance are correct and achievable ideals for external evaluators. He advocates that interviews should be avoided or minimised, that the evaluator should never talk to the programme staff nor look at programme rationale – this he defines as the method for producing 'goal-free' evaluation. The intent of goal-free evaluation is to assess and report the results of the program regardless of the aims or goals that it was designed to achieve. His stated objective is to provide '*validity, credibility, and comprehensibility of the evaluation'* (pg 483). He goes on to say that he views participatory or empowerment evaluation as 'sloppy'.

These two authors present what appear to be two extreme perspectives regarding the conduct of evaluations. In fact it is more like comparing apples and oranges as they approach evaluation theory from two entirely different perspectives even though they share space on the 'valuing' limb of the evaluation tree. The evaluation process described by Stake (1997) has an outcome of personal and organisational development or even survival. On the other hand Scriven's (1997) stated objective is the provision of a valid, credible and comprehensive evaluation. Given these totally different objectives it is no surprise that the methods used to achieve them are so dissimilar.

However, this multiplicity of approaches raises the issue of what criteria should be used to evaluate evaluations. The task of developing such methods has been ongoing over the past 35 years led by evaluators and professional organisations (Joint Committee on Standards for Educational Evaluation, 1994, Yarbrough et al., 2011). The next section provides the details of this process and the subsequent development of guidelines regarding the conduct of metaevaluation.

3.3 Metaevaluation

Introduction of the term metaevaluation' is attributed to Scriven (1969), who argued that the presentation of inaccurate or biased reports by evaluators could seriously mislead the public and encourage the adoption of products (in his case educational tools) that might be inappropriate or even detrimental in terms of impact, or the inappropriate use of funds for programmes that had no impact. He therefore advocated what he termed metaevaluation by which he meant '*any evaluation of an evaluation, evaluation system or evaluation device*' (pg 37). Scriven's definition provides us with a clear, if somewhat simplistic concept of how he envisioned the process could be used. That is, it could provide the framework for the critical appraisal of:

- an evaluation that has taken place
- a set of evaluation tools used to evaluate a system
- an evaluation device – such as a testing system

Following on from this Stufflebeam (Stufflebeam, 1974) provided a lengthy report that included an overview of the key evaluation issues that he believed needed to be addressed and the first proposals of how they could be managed. This paper forms the basis for the development of concepts of utility, feasibility, propriety and accuracy. It also set the stage for evaluation that is seen as collaboration between the evaluators and the stakeholders.

Subsequently, substantial effort was invested in more clearly defining terms and refining the process of conducting a metaevaluation. Metaevaluation is now seen as a professional obligation (Stufflebeam, 2001b) and this expanded definition of metaevaluation is now in common use: *Metaevaluation is the process of delineating, obtaining, and applying descriptive information and judgmental information about an evaluation's utility, feasibility, propriety, and accuracy and*

its systematic nature, competence, integrity/honesty, respectfulness, and social responsibility to guide the evaluation and publicly report its strengths and weaknesses (pg 183).

It is important to differentiate between metaevaluation and meta-analysis. Meta-analysis, the integration of findings from a number of different empirical studies, has its roots in the works of people such as Smith and Glass(Glass, 1976, Glass, 1977, Smith and Glass, 1977). The seminal work by Glass, McGraw and Smith (1981) provided the basis for the statistical analysis of data from a variety of studies and has been added to methodologically by statistical experts, especially the work of the statistical methods group of the Cochrane Collaboration(Cochrane Collaboration, 2011). This is very different however to metaevaluation where the purpose is to assess and critically examine the quality of a particular evaluation.

The activities recommended as part of a metaevaluation are presented in Table 4.

Table 4 Metaevaluation activities

Activities
1. *Determining and arranging to interact with the metaevaluation stakeholders*
2. *Establishing a metaevaluation team*
3. *Defining the metaevaluation questions*
4. *Agreeing the standards to judge the evaluation system*
5. *Negotiating the metaevaluation contract*
6. *Collecting and reviewing pertinent available information*
7. *Collecting new information as needed*
8. *Analysing the qualitative and quantitative information and judge the evaluations adherence to the selected evaluation standard.*
9. *Preparing and submit the final report*
10. *Helping the client and other stakeholders interpret and apply the findings*

Adapted from Stufflebeam (2000b)

3.3.1 Examples of uses of metaevaluation

Given this broadened definition, metaevaluation has been used in a wide variety of ways. The following is not comprehensive but provides examples of how the metaevaluation process has been adapted and used.

In the Philippines it was used to determine whether processes used to evaluate teaching performance met the standards of good quality evaluation (Magno, 2009). The Office for the Coordination of Humanitarian Affairs (OCHA) used metaevaluation to identify recurrent findings, conclusions and recommendations and to assess the quality of management practice for follow-up to evaluations (Robert and Engelhardt, 2009). A German retrospective metaevaluation of an organic farm programme (Eichert, 2008) showed that it was not always possible to evaluate all the components of the German metaevaluation checklist (DeGEval, 2008) owing to the limitations within the project reports. In spite of this, the evaluation was rated very highly. Only 30 items were marked as impossible to evaluate, 45 were marked as unmet while 191 being marked as met. In Australia, Reynolds (2006) used metaevaluation techniques to inform the role of NGOs. In Denmark it has been used in the business sector (Danida's Evaluation Department, 2004).

3.3.2 Metaevaluation tool

As can be seen from these examples, metaevaluation has the potential to be used in a number of different settings for a number of different purposes. The Joint Committee Standards for Program Evaluation are the most well known (Joint Committee on Standards for Educational Evaluation, 1994) and have been adapted for use internationally. These standards were developed in 1981 in what appears to have been an attempt to professionalise evaluation practices (Shadish et al., 1991). They have been updated over time through extensive consensus processes that has had input from a wide range of programme evaluators and theorists. The most recent update spanned a ten year period and has just recently been released (Yarbrough et al., 2011). The changes in this update are outlined at the beginning of the results section of this thesis.

The programme evaluation standards are based on the four aspects of evaluation that have been identified as the central to the conduct of the evaluation. These are outlined in Table 5.

Table 5 Program evaluation standards

Standard	# of items	Criteria
Utility	7	Based on the extent to which the stakeholders find evaluation processes and products valuable in meeting their needs – that is the uses for the evaluation
Feasibility	3	Based on the logistical and administrative aspects of the conduct of the evaluation
Propriety	8	Based on what is proper, fair, legal, acceptable and just in the conduct of the evaluation
Accuracy	12	Based on the truthfulness of the evaluation propositions and findings

Joint Committee on Standards for Educational Evaluation (1994)

The four standards are made up of numerous items each of which includes ten factors that require assessment. Stufflebeam (1999) developed a tool that includes the standards and provides formulae for the quantitative analysis and assessment of the quality of individual programme evaluations (see Appendix 2). This tool allows the reader to score a given evaluation according to each of the four standards and calculate an overall score as well as assigning a pass/fail designation.

Although the standards have developed over time and are accepted by consensus, no formal validation of the scoring system was identified. During the process of standard revision in 1994 a validation committee was established but its primary focus appears to have been on the validation of the process used for revision with very limited assessment of the applicability or validity of the standards (Gould et al., 1995).

In a recent PhD study Wingate (2009) addressed the issue of reproducibility in grades awarded by different assessors applying the same standards to a given set of evaluations. In her study she asked students, experienced evaluators and evaluation theorists to apply the criteria to a pre-defined group of evaluations. She reports a very high level of variance in their assessments, and concludes that this has serious implications for the use of standards to judge the quality of a given set of programme evaluations. Such high variability could be due to differences in the

experience of the assessors, but it could also be related to a lack of clarity in the accepted standards and raises issues related to their validity. Given my experience with the use of similar quality assessment tools which are used in the evaluation of medical research I would say that the differences were a combination of both these factors. However, as will be discussed later, the use of the tool in this thesis demonstrated the lack of clarity around a number of the items in the checklist.

The tool can of course also be used by programme evaluators as a checklist when planning, implementing and reporting their evaluations to ensure that all key domains are included. In this instance the tool could be used in much the same way as the CONSORT (Moher et al., 2001) or PRISMA (Liberati et al., 2009, Moher et al., 2009) tools which are used by journal editors to assess the reporting of randomised controlled trials and systematic reviews in the area of medicine.

It is interesting that in the development of metaevaluations in education, there are similarities to, as well as differences from the use of systematic reviews and meta-analysis in healthcare. Both became topics of public discussion in the mid 1970s (Madaus and Stufflebeam, 2000, Sackett et al., 1997). The similarities come from publications outlining the perceived lack of good quality research/evaluation, lack of methods to appraise such research/evaluation and most importantly a lack of knowledge of how to implement the findings from good research/evaluation in practice. However, systematic reviews and meta-analysis in healthcare have focused on establishing the most effective care treatments while the more limited process of metaevaluation concentrates on the quality of the evaluation itself with no consideration of the outcome.

Hammersley (2002) argues that this is appropriate. He makes the case that the research evidence in education is different in character from that provided in medicine and therefore the model used in evidence-based health care will not fit education. He notes that in medicine the recommended treatments are primarily provided to individual patients, while in education the research findings are usually applied to groups of students. On the other hand, Nutley et al (2008)

outline the similarities of implementation issues in education and medicine and call for more effort to integrate the findings from research studies into practice.

Reese (1999) points out that although lip service is given to the application of education research evidence in practice, there is limited evidence to demonstrate that this policy has been implemented. He outlines the history of educational research and argues that its quality remains problematic. This is a point also made by Lagemann (Langermann, 1989, Langermann, 1997) in two separate publications.

Be that as it may, the way forward has been different. Regardless of these limitations, metaevaluation is currently being used to assess programme evaluations, and it has been used as one aspect of this research report in an effort to assess the quality of the evaluations that were carried out during the lifetime of the SEARCH Program.

However, there may be other lenses through which to view this inspection of evaluations and evaluation processes. Two such lenses were considered: realistic evaluation and developmental evaluation. These are briefly discussed here.

3.4 Realistic evaluation

Realistic evaluation was introduced by Pawson and Tilly (1997) and has been described as a new paradigm in evaluation research based on a scientific *realist* approach. The focus is on the identification of problems within existing programmes. The evaluator's role begins with the identification of the programme's mechanisms, context and outcomes which assist in the development of ideas about what might work, for which group or individual and in what circumstances. The evaluators then take on the task of multi-method data collection and analysis to inform the development of, or changes to, programme specifications. These activities are carried out as part of an iterative process.

The concept of detailed examination of the mechanisms, contexts and outcomes in programmes is initially appealing. However, the realistic approach implies that there is a right or correct way to identify these three elements and that there is also a correct way to assess the data and implement the changes. The context in which

the SEARCH Program functioned was complex and evolving. Consequently what might be considered the correct approach for one cohort of students could well have changed in the next, that is it was unlikely that there was a definitive 'correct' approach, and there was a need for rapid assessment and evolution during the early iterations of the programme. Therefore the realistic evaluation lens was not selected for consideration in this thesis.

3.5 Developmental evaluation

3.5.1 Developmental evaluation – history and purpose

Developmental evaluation was introduced by Patton (2008) in his discussion of the concepts of utilisation focused research in the late 1980s. However, the world of evaluation continued to evolve and Patton himself has worked with numerous stakeholders in defining the concepts of developmental evaluation since that time. The first publication dedicated to developmental evaluation was made available following a two year iterative process that was the result of a number of workshops with voluntary organisation in Canada. The publication is the *Developmental Evaluation Primer* (Gamble, 2008).

In his subsequent book Patton (2011) describes the stages of his thinking in relation to developmental evaluation. These thoughts evolved from situations in which formative or summative evaluation did not fit the needs of the programmes being evaluated. His specific example relates to a project where he was contracted over a five-year period to evaluate an innovative community leadership programme. During the first two years the programme went through formative evaluation and substantive changes were made. However, when Patton announced at the beginning of the third year that the programme would no longer be allowed to change because they were entering a phase of summative evaluation, the programme staff were, to say the least, not best pleased. They saw the value of their programme as its ability to continually adapt to the needs of the community – they had no desire to implement a fixed programme model for the purposes of evaluation. Developmental evaluation was officially born. I say officially as it is almost certain that the concepts of developmental evaluation had been forming for

a number of years in the minds of Patton and others, but he has identified this as a defining moment in that evolution.

A clear definition of developmental evaluation is somewhat elusive. The following definition is provided by Patton (2011) when he writes that '*it guides action and adaptation in innovative initiatives facing high uncertainty'* (pg36). As described by Gamble (2008), developmental evaluation is most appropriately used in situations where there is high complexity and the innovations are taking place in a new or early stage of social innovation. The application of developmental evaluation is therefore limited and is described as a process to support innovation within evolving programmes and institutions. Patton's (2011) proposed purposes and uses of developmental evaluation are outlined in Table 6.

Table 6 Purposes and uses of developmental evaluation

Purpose	**Use**
Ongoing development	To adapt an innovative initiative to new conditions in complex dynamic systems
Adapting effective general principles	The use of ideas or innovations taken from elsewhere to be developed in a new setting
Developing a rapid response	In cases of major change or crisis used to explore real-time solutions and innovations
Performative development of potentially scalable innovation	The use of evaluation to bring innovative programs to the stage they are ready for formative or summative evaluation
Major systems change and cross-scale developmental evaluation	Providing feedback regarding the evolution of major change and how this might impact on the broader dissemination of a project (horizontal and vertical scaling)

Adapted from Patton (2011) pg 21-22

3.5.2 Developmental evaluation versus traditional evaluation

One could argue that developmental evaluation is really no different from traditional summative or formative evaluation where evaluation techniques are used and then changes are made to the target programmes or institutions. Patton (2011) contends that there are significant differences. He argues that there are seven domains in which differences can be identified. Importantly, he cautions that his comparisons are made on the understanding that there are numerous different types of evaluations and he is comparing developmental evaluation to the overall concept of evaluation including both summative and formative evaluation.

In the first domain of purpose and situation he envisions traditional evaluations being conducted to improve or validate existing programmes within relatively stable environments with the aim of finding out whether the programmes work. Developmental evaluation, he argues, is designed to support the development of innovations in complex and dynamic environments with the primary purpose of exploring possibilities and experimenting with innovations without the goal of arriving at a fixed intervention. He depicts situations where developmental evaluation would be used as those where the approach to implementation is 'ready, fire, aim' as opposed to the standard programme implementation of 'ready, aim, fire', an approach taken in from management literature and promoted by Peters and Waterman (1982).

In terms of the target he describes developmental evaluation methods as looking at system change in order to provide timely feedback, so that innovators can make sense of what is happening as top-down and bottom-up forces meet. By contrast, traditional evaluation is outcome focused and designed to fine tune existing and frequently static systems.

The focus of methods used in developmental evaluation is utilisation, the thinking is system-based with an emphasis on collaboration between participants and evaluators while they identify both the expected and the unexpected. By contrast, traditional evaluation has a linear (cause and effect) evaluator-established basis that attempts to rigorously measure performance (or lack of it) and apply deductive reasoning.

The roles and relationships in traditional evaluation can vary significantly. However, Patton (2011) describes the traditional role of the evaluator as independent, accountably focused outward (toward external authorities) with functions delegated to the organisation. He sees the developmental evaluation evaluator as a collaborator, and facilitator whose purpose is to introduce concepts of evaluative thinking. Accountability is to the programme or institution being evaluated with a focus on the programme and the environment in which it is situated.

The results of traditional evaluation are frequently validation and dissemination of what has been determined to be best practice. In his discussion of this facet, Patton (2011) also portrays a belief that traditional evaluation frequently engenders feelings related to 'fear of failure' and the result is often a detailed and ponderous evaluation report. By contrast, developmental evaluation, he asserts, is designed to nurture the participants and the reports are in the form of rapid real-time feedback that can be used immediately.

His penultimate point compares the views of complexity taken in the two approaches. In traditional evaluation, he sees the evaluator's as controlling the design, implementation and outcome of the process within the context of predictability and certainty. In developmental evaluation on the other hand, the evaluator expects uncertainty and lack of predictability and, from this perspective, there is a need to remain mindful of the evolution of the programme being evaluated and respond to those changes.

It is not the purpose of this review to delve into concepts of complexity. However, complex situations are defined as those where there is nonlinearity, emergence of patterns and dynamic interaction between subsystems (Goldstein, 2008). The definition goes on to describe the situation as adaptive when there is uncertainty and co-evolution between the agents involved (Patton, 2011). This is certainly the context in which developmental evaluation theory is described.

Finally Patton (2011) portrays the skills required by evaluators within each paradigm. He acknowledges the need for competence, experience and adherence to accepted standards in both camps. However, he emphasises that evaluators utilising developmental evaluation exhibit flexibility, adaptability, critical thinking and especially the ability to work with teams. Informed action must follow appropriate reflection.

In his primer on developmental evaluation, Gamble (2008) provides a list of questions (and rationale for those questions) that need to be addressed if one is considering the use of a developmental evaluation approach (Table 7).

Table 7 Appropriate space for application of developmental evaluation

Question	Rationale
What is driving the innovation?	Developmental evaluation is particularly appropriate if an organisation expects to develop and modify a program over the long term because of constantly shifting needs and/or contexts. It is helpful to distingquish between innovation taking place within an organisation and the adoption of an external innovation, which may not need developmental evaluation.
Are the proposed changes and innovations aimed at deep and sustained change?	Developmental evaluation is aimed at innovations that are driving towards transformational changes. Organisations often fine-tune their programs, and having an evaluative lens on those changes can be helpful; however the intensity of developmental evaluation may not be warranted in every instance.
Do we have a collaborative relationship with another organisation in which there is innovative potential in combining our respective talents?	Developmental evaluation may help different organisations work together through the effort to innovate. In this situation, the developmental evaluator can help the organisations through some of the inevitable tensions of collaborating and can provide a measure of transparency about the experiment.
Under what conditions does the organisation currently innovate? Is innovation part of the culture of the organisation	If this is already part of the culture, then the developmental evaluation role might be one that people within the team already play. If there is not a culture of innovation but there is a commitment to build one, then developmental evaluation might be helpful in stimulating that.
What are some core elements of what we do that we don't want to change?	There might be elements of an initiative that are known to work, or for another reason are expected to stay the same. Evaluation requires resources, and if things will not change, these resources are better directed elsewhere. If something is not going to be adapted but there is interest in finding out whether it works or not, a summative evaluation is appropriate.
Is it clear for whom the evaluation is intended?	This is a vital question for any evaluation, developmental or otherwise. For an organisation to make good use of developmental evaluation, it is important to have key decision makers interested in and open to using evaluative feedback to shape future actions. If the only user of the evaluation is external to the innovating team (such as a funder), then developmental evaluation is probably not the appropriate approach.

Gamble (2008)

Given this background to developmental evaluation and the knowledge of the processes used in the evaluation of the SEARCH Program it was decided that developmental evaluation would be the second lens through which those processes and responses would be examined.

3.6 Conclusion

This section has provided an overview of programme evaluation, the current theoretical positioning of the well known theorists as well as a brief look at the history of evaluation practice. It has looked in depth at the practice of metaevaluation and developmental evaluation which form the basis for the research reported in the remainder of this thesis.

Combining the historical perspective of evaluation and the context of the SEARCH Program it became clear that there would be value in examining the evaluations from more than one perspective. Given the pragmatic approach used in the development of the SEARCH Program there was no question that approaches from the 'use' branch of the evaluation theory tree should be used. Having made that decision, deciding to use an existing quantitative metaevaluation analysis tool seemed a logical next step. The extensive evaluations carried out and the changes to the SEARCH Program during its existence meant that the broader examination of the role of those evaluations in programme development could prove to be informative and allow for exploration of the role of developmental evaluation. How these two decisions were implemented is presented in the next chapter.

4 RATIONALE AND METHODS

This chapter provides an outline of the research methods used in the research that formed the basis of this thesis. Included are descriptions of the methods of data collection and the data analysis procedures employed. The chapter concludes with a rationale for the use of a case study approach.

The research questions presented earlier are addressed through a retrospective case study. Two approaches have been used. The first is a metaevaluation that included the quantitative assessment of the evaluations conducted during the life of the SEARCH Program. The second is a qualitative analysis informed by these evaluations and SEARCH Program documents.

4.1 Metaevaluation

Over many years, Stufflebeam and others have formulated the aims and key activities that constitute the conduct of metaevaluation (Joint Committee on Standards for Educational Evaluation, 1994, Stufflebeam, 1974, Stufflebeam, 2000b, Stufflebeam, 2001b). The purpose of metaevaluation is the assessment of the content and quality of reported programme evaluations. Currently ten activities are included in recommendations for the conduct of the metaevaluation process which were reviewed in Chapter 3 (Stufflebeam, 2000b).

These activities are presented here as occurring in a linear fashion and in some cases that is how they occurred. However, there was also an iterative element to the process of examining SEARCH Program data that included re-visiting decisions taken in the early stages of the metaevaluation process. The following section outlines the ten required metaevaluation activities and how they were addressed in this project.

1. *Determining and arranging interactions with the metaevaluation stakeholders*

My previous involvement in the SEARCH Program meant that I had observed and participated in programme delivery on a number of occasions, attended SEARCH conferences and had extensive contact with the SEARCH Program faculty. This included hosting a conference, sponsored by the Strategic Health Authority, in

Liverpool in 2002. The conference was attended by four SEARCH faculty and one SEARCH participant. The focus of the conference was to introduce the SEARCH Program model to a UK audience and determine whether there would be interest and support for a UK based programme.

Specific discussions related to the research reported in the present thesis took place over a period of 18 months, from January 2008 until June 2009. Discussions were in person (in Alberta) and through conference calls and email communications. In February and June of 2009, I spent a total of 10 days in Alberta meeting with SEARCH faculty members and participants to finalise the scope of this research project. Discussions with these stakeholders focused on how the SEARCH Program had been evaluated and it was agreed that my work on the project would focus on these evaluations. As noted earlier, the programme was unexpectedly terminated because of budget cuts in June 2009. As a result, although informal communication is still taking place with former faculty members of the SEARCH Program they did not participate formally as the project proceeded.

2. *Establishing a metaevaluation team*

Given the nature of the educational requirements for the PhD process, I am the team. However, I sought and received input from former SEARCH Program faculty members, external peer reviewers and my PhD supervisors.

3. *Defining the metaevaluation questions*

This activity was undertaken in consultation with former SEARCH Program faculty members, including the former Director, a core module leader and a programme administrator, and my PhD supervisors. The metaevaluation questions to be addressed are reflected in the research questions that directed this research and include an assessment of the quality of the evaluations conducted during the SEARCH Program.

4. *Agreeing the standards for judging the evaluation system*

The standards for evaluation as developed by the Joint Committee on Standards for Educational Evaluation (Joint Committee on Standards for Educational

Evaluation, 1994) were used to judge the evaluations that had been carried out as part of the SEARCH Program. In particular, the tools and analysis formulae provided by Stufflebeam (1999) were used to calculate the quality scores for the evaluations. (see Appendix 2)

5. *Negotiating the metaevaluation contract*

In early June 2009 a draft memorandum of understanding between myself and the then Director of the SEARCH Program was prepared. The agreement included consideration of protocol development, ethics approval, anonymity of data and the scope of evaluations to be included and report timelines. The premature termination of the programme meant that this memorandum of understanding was not ratified. However, the project has proceeded in a manner consistent with the principles agreed at that time.

6. *Colleting and reviewing pertinent available information*

In February 2009 the SEARCH co-ordinator made available a comprehensive master list and where available electronic copies of all programme evaluations conducted during the operation of the SEARCH Program. These evaluations included interim and final cohort evaluations as well as evaluations related to participating organisations. In addition, in June 2009 access was provided to all electronic files related to the delivery and evaluation of the SEARCH Program. These files included evaluation reports, correspondence, and contractual information. Also included were minutes of faculty and evaluation strategy committee meetings that dealt with the planning or results of evaluations and any planning documents that detailed the faculty response to evaluations. There was a limited amount of data related to the financing of the programme; in particular, detailed information related to the external contracting of evaluations was not available.

7. *Collecting new information as needed*

No new data beyond the electronic data provided were gathered.

8. *Analysing the qualitative and quantitative information and judging the evaluation's adherence to the selected evaluation standard.*

For the purposes of the present research, reports and evaluations (from the period of 2000 to 2005) included on the master list were placed in categories. These were cohort specific evaluations, general evaluations, an Information Technology (IT) action research project, evaluation frameworks and other general reports such as AHFMR general reports, SEARCH Canada and SEARCH Light reports.

Evaluation reports (cohort specific and general evaluations) were read and narrative data extracted to a bespoke EXCEL® database. Data from three reports were extracted, the database was re-examined and changes were made to reflect more clearly the required data extraction and, as far as possible, provide consistency in data extraction. Data from the remainder of the evaluations were then extracted. Data from specific cohort studies, conducted by a single evaluator over a period of time were grouped together for data extraction and analysis (e.g. cohort evaluations using multiple methods over an extended period of time). The grouping of these evaluations is clearly outlined in the results section.

Descriptive data extraction included basic details of the evaluations (e.g. cohort, date) as well as data categorised using the RUFDATA categories (Table 8) and the evaluation focus on the level of evaluation (Table 9). RUFDATA as outlined by Saunders (2000), is an acronym that identifies categories used to provide structure for decision-making to shape evaluation activities. In this case they were used to provide consistency in data extraction from evaluations that had been undertaken in the programme evaluations.

Table 8 RUFDATA categories and definitions

Category	**Definition – general**	**Definition – within this project**
Reasons and purposes	Why is the evaluation being done?	Evaluation aims/objectives
Uses	How will the evaluation results be used?	Programme development, strategic planning
Foci	What activities will be evaluated?	Programme, experience, impact
Data and evidence	What will be evaluated?	Data collected
Audience	Who will use the evaluation?	SEARCH administrators, faculty, individuals, participating organisations
Timing	When should the evaluation take place?	Timing of evaluation in relation to module delivery and programme history
Agency	Who should conduct the evaluation?	Defined as internal, or external

Saunders (2000)

Assessments of the levels of evaluation outcomes also outlined by Saunders (2007) were designed to assist in the organisation of programme or policy evaluations and are used here to provide consistency in data extraction and to allow evaluations to be compared. These levels come from a work-based learning perspective and are informative because the SEARCH Program provided a direct link with the work environments of both participants and faculty.

Table 9 Levels of evaluation focus

	Definition
Level 1	Quality of the experience of the SEARCH Program from the perspective of participants, faculty or health authority
Level 2	Quality of the situated learning outcomes including skills acquisition and knowledge acquisition
Level 3	Quality of transfer or reconstruction of learning to the work environment
Level 4	Quality of organisational impact
Level 5	Impact on macro or long term strategic objectives (individual, SEARCH or participating organisation

Saunders (2007)

Quantitative data were extracted using the standards for evaluation developed by the Joint Committee on Standards for Educational Evaluation (1994, Stufflebeam, 1999). A summary is presented in Appendix 2. Data were extracted into a bespoke ACCESS® database. These data were then exported into a second bespoke EXCEL® database for appropriate data analysis as set out in the standards protocol and using the formulae provided by Stufflebeam (Stufflebeam, 1999).

In acknowledgment of the possible bias exerted through data extraction conducted by a single reviewer, as well as possible data extraction errors, a data cross-checking mechanism was piloted. Quantitative data from three evaluations were extracted by one reviewer, on two separate occasions. Results were compared and where the assignment of the scores did not match the items were re-examined and a final decision made. On average there were 12 discrepancies over the 300 data points (range 10-14) in each of the three pilot evaluations and therefore the process was not repeated for the remainder of the evaluations.

9. Preparing and submitting the final report

This thesis represents the final report of this analysis.

10. Helping the client and other stakeholders interpret and apply the findings

Given the termination of the SEARCH Program it is not possible to apply the findings directly to that programme. However, it will be possible to identify implications related to the development and evaluation of CPD programmes for healthcare professionals.

4.2 Developmental evaluation

Qualitative data extraction related to developmental evaluation was carried out after the quantitative metaevaluation had been completed. As noted earlier, electronic records of the SEARCH Program were made available for this research project. The documentation used in the qualitative analysis included records of the SEARCH Steering Committee and the SEARCH Evaluation Committee meeting minutes and evaluation documents for the period of 2000 to 2005. The first of these dates was selected for pragmatic reasons – there were no identified electronic records identified prior to this time. The second date represents the time point when the SEARCH Program was no longer a part of AHFMR and therefore had a different structure. It is worth noting that the evaluation of the SEARCH Program continued until 2009, however those evaluations are not considered in this research project.

All available minutes and the documents for the meetings over this six year period were first read to provide an overview of the evolution of the programme. The

documents were then examined more closely and data extracted using Atlas.ti®, a software package designed to organise and analyse qualitative data.

A qualitative *directed thematic content* approach was used to extract and analyse the data in relation to aspects of developmental evaluation (Krippendorff and Bock, 2009). The initial purpose of this directed analysis was in the first instance to determine whether the programme was functioning in a complex and uncertain context. Or, in Patton's (2011) terms, did it fit within the definition of a complex programme functioning in a complex environment? The second purpose was to provide an overview of evaluative practices. That is, how were evaluations planned and conducted and more importantly how were the results used to make changes to the programme. This information was then used to determine whether evaluations had served a developmental role. This was done within the context of the concepts of developmental evaluation.

Anderson (2007) provides a rationale for the use of a thematic approach. She maintains that thematic content analysis identifies the common themes within the texts and is fundamental to all qualitative analysis. She goes on to argue that the analysis is therefore objective in nature. However, in this current research that is certainly a debatable point as the document selection and themes were both previously established by what was available and then through my grouping and selection. Given my history with the programme I would not label myself as objective. However, I have attempted to make the process transparent so that the areas of potential bias are identified. Data were therefore extracted into thematic categories within three themes that examined the environment in which the programmed functioned, the evaluative practices/culture used as the programme evolved and programme innovation. These categories are defined in the results section of this thesis.

Content analysis has a long history but made its modern debut as a quantitative research tool when, in 1910, when Weber advocated it as a means to critically examine the content of newspapers (Krippendorff and Bock, 2009). Its evolution into the qualitative paradigm was gradual (Zhang and Wildemuth, 2009) and happened as it was identified as a means by which not only a word count but an

analysis of the social context or social reality of the situation could be examined. In contrast to standard content analysis where the focus is on word count, qualitative content analysis is based on selected texts that assist in the investigation of the specific research question and in this sense is therefore directed.

It is important to consider whether such a process is inductive or deductive. Inductively it would first include examination of the raw data and the process of open coding to derive the categories – as used inductively in grounded theory (Glasser and Strauss, 1997). If it is deductive then it would employ pre-established and somewhat evolving codes relating to the research question and use the data to inform the development of a conceptual framework or theory. The first of these was definitely not used in this project but a case could be made that the second was applied. However, a closer fit is to what Krippendorff and Bock (2009) call abductive reasoning – that is neither inductive nor deductive but moving from one kind of incidence – that is readable text to inference regarding the environment. They go on to describe '*a model of the relationship between textual matter and the empirical domain of the desired inferences as an analytical construct'* (pg 105). Such constructs, they argue, provide a connection between the data and the context that is being examined and more closely fit what was done in this research.

4.3 Ethics

The question of ethical approval for secondary research is an interesting ethical dilemma. Had this research project been limited to the examination of the publically available evaluation reports then there is a question, similar to the one when systematic reviews of research literature are carried out, whether any approval is required. However, given that this research also extended to include examination of internally held documents then ethical approval and the maintaining the anonymity of the faculty and participants was important. Having said that – certainly the faculty members will likely be able to identify their own and their colleagues in the quotations used in the qualitative analysis.

Ethics approval was sought and received through Lancaster University. As noted above, ethical issues related only to the anonymity of the SEARCH Program faculty members and participants and these were managed through the anonymous nature of the evaluations. All specific references to faculty in the analysis of faculty minutes were omitted or replaced with impersonal coded letters.

4.4 Rationale for using a case study approach

The research aims were achieved using a case study approach. However, the term case study is used in different ways in the literature and in practice requires clarification and discussion in relation to this thesis.

It is worth highlighting the interchangeable use of the terms methodology and methods in relation to case study research. Klein (2007) points out how the term methodology should be used in the global study of research methods; it should be used only when examining the broader theoretical perspectives such as Popper's empirical falsification or Kuhn's paradigm shifts. By contrast, Klein (2007) defines that methods simply as the techniques used in performing the work.

However, in the context of a case study this is not a clear cut definition, as is demonstrated in this thesis. In the research reported here, the term 'case study' does not describe the methods used; instead, it describes the approach used, which incorporates a variety of of data collection and analysis approaches. This could then be seen as a mixed methods research with all the benefits and pitfalls described by Brannen (Brannen, 2005). However, in strict terms mixed methods research uses various data collection methods to answer the same question. This research uses various data collection and analysis methods to address a range of questions and therefore does not fit with this definition of mixed methods research.

The term 'case study' has been used in numerous ways over time and it is worth providing some of this background and presenting some of the controversy that surrounds its use (Byrne, 2009). The results of a comprehensive review of the history of case study methods (1900-1990) were published by Platt in 1992. She provides an excellent overview of the waxing and waning of the use of case

studies in research and of the various and differing definitions as well as the political impact on the use of case studies (e.g. pre- and post-war). She outlines the original use of case studies as seen in social work to refer to specific individual cases. She then provides a discussion of the evolution of case studies and their use in both individual and broader sense– e.g. the nation as a case.

Platt (1992) clearly demonstrates the discrepancies between case study methodological discussions of case studies provided by academic writers and the lack of case studies to serve as exemplars of those academic perceptions. That is, in methodological discussions, the case study is seen as the use of extensive data to examine all aspects of the individual case and therefore includes what is unique, as well as holistically examining all aspects of the case. She maintains that the majority of case studies were neither comprehensive nor holistic and therefore identifies a mis-match between what the academic literature says a case should be and how researchers reported them. Her excellent review ends with a listing of more recent texts published in the area and a critique of the first publication of the textbook by Yin who has come to be a leader in the use of case study methodology (Yin, 1984). This critique is discussed later in this section.

Hers however is not a consensus opinion. Tight (Tight, 2010) recently provided an overview of the topic. In his opinion, the term case study should not be used at all. He advocates 'a tell it like it is' approach and encourages researchers simply to state what they did -e.g. '*a detailed examination of....'*(pg 338) so as not to get caught up in using the ill-defined term of case study.

In her recent book, Simons (Simons, 2009) maintains that the '*primary purpose for undertaking a case study is to explore the particularity, the uniqueness of a single case'* (pg3). She takes this perspective within the concepts of naturalistic enquiry. Unfortunately, taken outside that context, this definition does case study research a disservice and perpetuates the concept that case studies are unique and therefore their findings cannot be transferred or generalised.

Flyvbjerg (Flyvbjerg, 2006) on the other hand provides an excellent overview and points out how staunch quantitative researchers (e.g. Campbell and Eysenck)

came to alter their views and value the contribution of case studies especially in the area of social science where predictive theory does not exist, and is unlikely to do so in the future. In his detailed paper Flyvbjerg (2006) goes on to present and refute five misunderstandings of the use of case studies (Table 10). His article is lengthy, and only a short summary is provided in the table which does not do justice to his extensive arguments.

Table 10 Misunderstandings of case study research

Misunderstandings	**Refutation**
General, theoretical (context-independent) knowledge is more valuable than concrete, practical (context-dependent) knowledge.	Concrete experience is necessary as distance from the object of the study and lack of feedback deter from learning from the situations. Knowledge of experts is made up of data from thousands of cases.
One cannot generalize on the basis of an individual case; therefore, the case study cannot contribute to scientific development.	This depends on the case that is selected and how the data is used.
The case study is most useful for generating hypotheses; that is, in the first stage of a total research process, whereas other methods are more suitable for hypotheses testing and theory building.	Case studies can be used for both generating and testing hypotheses – again it is dependent on the selection of the cases
The case study contains a bias towards verification, that is, a tendency to confirm the researcher's preconceived notions.	Case studies can be rigorously designed. The depth of the research method used often points out more alternative outcomes than expected by the researcher.
It is often difficult to summarize and develop general propositions and theories on the basis of specific case studies.	The descriptive nature of the case study allows for in-depth analysis of the 'thick' data that is lost if an author attempts to provide a too short or too structured report.

Adapted from (Flyvbjerg, 2006)

Hammersley (2010) asserted that although case studies have been used ideographically, they can also be used to demonstrate a case with intrinsic interest. In the context of this thesis, the case is an innovative CPD education programme that was extensively evaluated over an extended time period. Hammersley (2010) goes on to describe how transferable lessons can be learned and applied to a sample from a larger, finite population to which the case belongs. In this instance the case is used as an exemplar that will resonate within the wider context of CPD delivery.

Hammersley (2010) also outlined the key information that is required in order to make these generalisations. Important questions include: What is the population?

Why is it important? What are the units or cases? How were the cases selected? What justification can be provided for using evidence from the cases(s) to draw conclusions about the population? He adds an appropriate warning that such generalisations should not be used to comment on causal relations in these situations.

The need to be clear about the purpose of the research was supported by Trowler (2010) who made the point from a slightly different perspective. He stressed the importance of differentiating between the use of data to test/refine/develop theory and the use of theory to interrogate data either to provide an organisational structure/order for the data or to explain it from a theoretical standpoint. In the context of the research reported in this thesis, the data are used to assess the quality of the evaluations using an established metaevaluation tool and to explore, refine and contribute to the evolving discussions regarding the concepts of developmental evaluation.

In the late 1980s Ragin and Becker (1992) attempted to address the challenges of using cases in social research. Through a hosted symposium, case study experts (including Platt) addressed key topics designed to define what could be considered a case. There was general agreement that cases should be chosen for theoretical or purposive reasons and should not be selected randomly.

In this edited book, Ragin (1992b) provides a mapping of cases as depicted in Table 11. The table makes it seem quite easy to position my case – the SEARCH Program is both an empirical unit and a specific case that has been identified. This is important as there is one area of consensus in all publications about the use of case studies, that it is critical to have a well-defined case.

Table 11 Conceptual map for cases

Understanding of cases	**Case conceptions**	
	Specific	**General**
As empirical units	1. Cases are found	2. Cases are objects
As theoretical constructs	3. Cases are made	4. Cases are conventions

Adapted from Ragin(1992b)

In this research project the SEARCH Program is considered the case. Within this case a number of different data sources are identified including programme evaluations, faculty meeting minutes, evaluation steering group minutes and programme planning documents. Gerring (2007) takes the view that these documents should be treated as within case observations, and that was how they were viewed in the research project reported here. However, Ragin (1992a) also discusses what he calls 'casing' which allows iteration as the case is defined. This is discussed in more detail in the discussion section of this thesis.

Yin (1984) has been recognised as a leader in the area of case studies and the first edition of his book on case studies in 1984 has been credited by Platt with having raised awareness and respect for the use of case-study evaluation (Platt, 1992). In the updated and clearly written fourth edition of his book, Yin (2009) explains that case studies can be used to accomplish four different goals. They are:

- *to explain – causal links that may be complex and not lend themselves to standard research methods*
- *to describe – an intervention in the real life context*
- *to illustrate – a specific area within an evaluation*
- *to enlighten – in situations where interventions may not have a clear or single set of outcomes*. (pg 19)

The research reported in this thesis is most closely linked to the description and enlightenment aspects.

Yin (2009) provides further explanation of the case study approach. He emphasises the need to formulate a clearly focused set of research questions, pointing out that these questions should ask how and why about events over which the evaluator has little or no control. He goes on to point out that the case-study inquiry:

- *Copes with the technically distinctive situation in which there will be many more variables of interest than data points, and as a result*
- *Relies on multiple sources of evidence, with data needing to converge in a triangulating fashion and as another result*
- *Benefits from the prior development of theoretical propositions to guide data collection and analysis* ' (pg 18)

He provides information on planning, protocol development, pilot testing and data analysis. He points out that case studies are perceived (wrongly) as an easy option when in fact they require intelligent investigators that have the ability to question constantly as the data are collected and to adjust data collection appropriately. He contrasts this to experimental studies in which the data collection instruments are set in advance and require limited intellectual input by data collectors as the data are being collected.

As an interesting link to evaluation, Yin was invited by the American Evaluation Association to reflect on the use of case studies in evaluation (Yin, 2000). In this paper he discusses case-study tools and emphasises the need for appropriate protocols to direct case-study evaluation. He goes on to identify three features that make up the profile of a case study. The first is that the case study is dependent on the use and integration of information from multiple sources that may be direct inputs, observations, interviews, documents or archives. He maintains that the conclusions for the case study need to be substantiated through the consistency of the data from the various sources accessed. Secondly, the methodology has to assume that there is a richness in the data that allows the researcher to examine a real life scenario. Thirdly, he goes on to say that the case study may be restricted to a single case or draw on data from multiple-case studies. He goes on to explain that the ability to generalise from the results of a case study *'depends on the development, testing, and replication of theoretical propositions (analytic generalization) – rather than any notions based on the selection of numeric samples and extrapolating to a population (statistical generalization)'* (pg186).

As noted earlier, Platt (1992) criticises Yin's approach to case studies, pointing out that it differs from the historical approach to published case studies. I would argue that Yin's approach closely fulfils the academic methodological aspects of case studies as outlined by Platt (1992) in that his approach fulfils her identified criteria such as the requirement to treat holistically, sets of specific data relating to one or more unique individual cases.

Given this background it was considered appropriate to use a case-study methodology to address the research aims of this project. Using terms from Yin

(2009) outlined earlier, the aims are *descriptive* in nature – that is, using metaevaluation it will examine the quality of the completed evaluations. However there is also a need for *enlightenment* –to determine whether these evaluations were used to inform programme development. To do this it will be determined if the evaluations meet the criteria set by developmental evaluation theories (e.g. developmental evaluation) and how can we use an examination of the evaluative practices used in the SEARCH Program to inform the growing field of CPD for healthcare professionals.

5 RESULTS

The SEARCH Program was the focus of extensive and continual short (module) and long term (up to 5 years) evaluation. These evaluations were guided through two different evaluation plans. The first was established at the inception of the programme and spanned the period from 1996 to 2001. Copies of this report were not available in either paper or electronic copy. The second was called the 'Evaluation Blueprint' and was the result of a collaboration of SEARCH management and faculty, an external consultant and invited facilitators. It provided the evaluation plan for the following 15 years. The initial documentation was the result of a retreat where through presentations and group discussions the model for future evaluations was drafted. The report then went through a number of iterations and became the guide for future evaluations.

This chapter presents the results of the quantitative metaevaluation analysis that was carried out using the metaevaluation tool. It then goes on to present the findings of the qualitative analysis of the SEARCH documents to demonstrate how these evaluations contributed to programme development. The qualitative analysis is structured within the concepts of developmental evaluation as presented by Patton (2011).

5.1 Metaevaluation standards changes

The literature review outlines the evolution of the evaluation standards recommended by the Joint Committee on Standards for Educational Evaluation (1994). It is these standards and the subsequent checklist developed by Stufflebeam (1999) that have been used as the basis for the quantitative metaevaluation of the SEARCH Program evaluations presented in this thesis. The Joint Committee on Standards for Educational Evaluation has recently published the results of a ten year consensus process that resulted in a revised version of the standards (Yarbrough et al., 2011). However, the data extraction for the quantitative component of the present research was almost complete when the revisions were released and a revised checklist and quantitative assessment tool was not available to accompany the new standards.

It was therefore decided to continue with the original data analysis plan. It is however worth noting the differences between the versions. A table of the versions of the standards is provided in Appendix 3. In summary, the four categories (Utility, Feasibility, Propriety and Accuracy) have been augmented with a fifth (Evaluation accountability) that is specific to metaevaluation. The items in this final category had previously been included in the Accuracy domain. The specific important differences are outlined in Table 12.

In general the changes are not substantive and in fact incorporate a number of the details that are outlined in the Stufflebeam (1999) checklist. It is unlikely that using the revised standards would have substantively altered the metaevaluation presented in this thesis except to limit the analysis, owing to the lack of quantitative analysis framework that is provided by the Stufflebeam checklist.

Table 12 Key differences between 1994 and 2011 Evaluation Standards

Domain code	Differences note
Utility U1 and 2	The first utility criterion is now related to the evaluator, and the stakeholder description is in the second slot
U3	There is explicit instruction regarding the need to continually negotiate the purposes of the evaluation based on the needs of the stakeholders
U4	Values are now required to be specifically clarified
U5	Now split into two new areas U5 and U6. The emphasis on a clear report remains but an additional mandate for description and the promotion of use of the report has been added
U6*	Provides more emphasis on providing reports as needed by the stakeholders as opposed to the previous version's emphasis on interim findings.
U8	The previous point provided an emphasis on the encouragement of follow-through and use while the new criteria have a focus on guarding against unintended negative consequences and misuse
Feasibility F1	The term effectiveness is introduced to replace the previous description regarding the use of practical procedures that limit disruption
F2	The previous description of differing perspectives has been simplified in terms of responsiveness and practicality. Notions of 'politics' have been included in a new F3 which covers the balancing of political needs
F3	Cost effectiveness is now termed 'effective and efficient use of resources'. It is worth noting that none of the previous or current documentation defines what they is meant by cost effectiveness
Propriety P4	The old P4 that related to respect during interactions has been deleted and has not been replaced.
Accuracy A1	The focus has been changed from a focus on clear description of the programme to the previous number 10 standard that had a focus on the justification of the conclusions and decisions
A2	Validity of information has moved up from the number 5 slot to the second slot
A3	Reliability of information has moved up from the number 4 to the number 3 slot
A4	As a result of these two previous changes, the program description and context have moved into the number 4 slot
A7	The previous two standards A8 and A9 which dealt separately with the issues of qualitative and quantitative data analysis have been combined in a more generic standard that talks about 'explicit evaluation reasoning'.
A8	Now the final standard in this category encompasses the need for scope that will 'guard against misconceptions, biases, distortions and errors.'
Evaluation Accountability	This is a new standard with three points that replace the previous A12, which recommended metaevaluation of all evaluations

*It is interesting that explicit mention of dissemination has been removed.

5.2 SEARCH evaluations

The summary list of evaluations provided from the SEARCH Program executive included 42 reports. Three were duplicated and two were reviews of capacity within the Regional Health Authorities (RHA) and were not directly related to the

SEARCH Program and were therefore excluded. An additional four reports were external to the SEARCH Program (e.g. evaluations in other provinces); none of these directly examined the SEARCH Program and they have not been included in this report or analysis. Overall 33 reports were therefore available for the analysis.

The reports were divided into five categories as shown in Table 13. Cohort specific evaluations included multiple reports. Where there were multiple reports from the same evaluator, which were presented in a global final report, they were classified for this analysis as one report. An exception was made for the SEARCH IV focus group report which was extensive and reported separately; in this case the descriptive data are reported separately but the SEARCH IV report is included only as a single entry in the quantitative analysis. Therefore, data from the cohort reports (6) and the general evaluations (5) were extracted into descriptive data tables as previously outlined using the RUFDATA and Impact Level frameworks and were also used in the metaevaluation standards analysis. In total there were 11 reports that were summarised for the initial narrative data extraction and 10 for the statistical analysis.

The three reports from the IT action research project did not lend themselves to quantitative analysis and are discussed separately as part of the qualitative analysis related to developmental evaluation.

Data were not extracted from the Evaluation Framework documents but they informed the qualitative data analysis. As noted earlier, the initial 1996-1998 evaluation framework was not available in either hard copy or electronic version. The second evaluation framework 2001-2005 included four reports. The remaining five reports were not specific evaluations of the SEARCH Program and their data were not extracted. However, the reports were read and used to inform the qualitative analysis.

Regularly scheduled module delivery provided evaluators with face-to-face access to participants and faculty and allowed for the conduct of structured interviews and focus group discussions. At other times evaluators travelled to specific regions to collect data or used telephone conferencing facilities.

In addition, extensive intra-module evaluations were conducted through feedback following module sessions. Summaries of these were not included in the overall programme evaluation list nor were they subjected to quantitative analysis. Consideration of these evaluation practices is however considered in the qualitative analysis of this report and in the discussions related to developmental evaluation.

Table 13 Summary of report categories

Report category	# of individual reports	Data	# in quantitative analysis
Cohort specific reports: SEARCH I SEARCH II SEARCH I and II SEARCH III **SEARCH IV summative **SEARCH IV focus group	 3 7 1 3 1 1	Interim reports assessed with summary report Data summary tables Included in descriptive and statistical analysis	6
General evaluations: Faculty Impact and Experience Project Tracking Organisational Impact Managers Survey Collaborative Network Evaluation	5	Data summary tables Included in descriptive and quantitative analysis	5
IT Action research reports	3	Used in qualitative analysis	0
Evaluation Framework reports	4	Used in qualitative analysis	0
SEARCH related: AHFMR general reports SEARCH Canada Expert Panel Review *SEARCH Light evaluations	 2 1 2	 Not directly related to SEARCH Program Related to SEARCH Canada not SEARCH Program Not directly related to SEARCH Program	0
Total number of reports	33		11

*SEARCH Light was the electronic newsletter of the SEARCH Program
** Quantitative standards data entered as a single evaluation

5.3 RUFDATA results

As outlined in the methods section data were extracted using the RUFDATA framework. The previously presented table is repeated here in Table 14 to allow for easy reference to the categories.

Table 14 RUFDATA categories and definitions

Category	Definition – general
Reasons and purposes	Why is the evaluation being done?
Uses	How will the evaluation results be used?
Foci	What activities will be evaluated?
Data and evidence	What will be evaluated?
Audience	Who will use the evaluation?
Timing	When should the evaluation take place?
Agency	Who should conduct the evaluation?

Saunders (2000)

Data extracted from the evaluations are presented in Table 15. The primary objectives/reasons for conducting the evaluations can be divided into three categories;

- To assess accomplishments by comparison with programme objectives
- To inform programme development
- To assess impact on participants, health regions and faculty in both the short and long term.

Early evaluations appear to have had a central focus on informing programme development while later evaluation objectives and methods began to examine the impact of the programme.

The first two of these objectives match the reasons for evaluation as outlined by Chelimsky (1997) earlier in this thesis. That is they look at accountability (is the programme doing what it set out to do) and the use of results to inform programme development. However, Chelimsky's (1997) third reason for evaluation relates to the acquisition of knowledge and understanding. A case could be made that assessing impact would fit into this category but I believe that such an interpretation is pushing the boundaries. Certainly the evaluation of the impact of a programme will examine knowledge changes, but in the context of the present study it has a more direct link to changes in practice, which are related to the participants, the health authorities in which they worked, and the faculty.

The uses of the evaluations were clearly defined in eight out of ten of the evaluations included in this project. As noted above, it was evident that the early evaluations were used to inform the ongoing development of the SEARCH

Program and programme modifications. In addition there was an emphasis on the impact of the programme on participants, faculty and the health regions.

The data relating to foci were somewhat difficult to extract. This is partially because the paper by Saunders (2000), which was the basis for this analysis, equates foci with the range of activities that could be evaluated. For the purpose of this current analysis foci were defined not as the range of activities but as the range of stakeholders that were the targets of the evaluation. These were clearly defined within the evaluations and included participants, faculty and health authorities. This is an adaptation of the RUFDATA category that allowed stakeholders to be identified as playing an important part in the evaluation process.

Data and methods of data collection varied but included analysis of both quantitative and qualitative data. Computer technology allowed for rapid electronic access to participants, faculty and in some cases stakeholders in the health regions. This was used to regularly collect feedback during residential modules and also to collect other survey data outside the module delivery periods and as part of the longer term follow-up evaluations.

Timing of the evaluations provided an opportunity for both formative and summative evaluation. As noted earlier cohort evaluations spanned a number of time periods including the period when the programme was running and the time taken for both short and long term follow-ups. In addition these were well planned evaluations with consistency of data collectors across the timeframe of the evaluations.

In terms of agency, all but two of the eleven evaluations were carried out by externally contracted evaluation consultants. Having said that, the two primary consultant groups were contracted on a number of different occasions, allowing them to become very familiar with the SEARCH Program and use insights from previous evaluations to inform the development of later evaluation activities. It also allowed the evaluators to build a rapport with SEARCH participants, faculty

and health authority stakeholders over time. Therefore their status as external or fully independent evaluators could be questioned.

5.3.1 Value of using the RUFDATA framework

The RUFDATA framework was designed to assist in procedural decisions related to programme evaluation and policy. It has been used here as one of a number of tools to retrospectively examine an evaluation process. The use of the RUFDATA framework in this way provided consistency in consideration of the various aspects of the SEARCH Program evaluations. The resultant data table provides a clear overview of the extensive evaluation activities that were carried out over the span of the programme. The data also demonstrate the integration of evaluation activities through the use of a limited number of external consultants who worked with SEARCH Program faculty to integrate knowledge from previous evaluation activities.

As noted above, in the case of foci it was somewhat difficult to match the outcome extracted with that defined by the author of RUFDATA. Consideration of this aspect allows for reflection on the other areas of the framework and demonstrates that although a definition was provided for each category the definitions were not so constricting as to make the process a ‘box ticking’ exercise but in fact provided enough structure to allow for exploration of issues while not confining the extraction of the data or the subsequent analysis.

In summary the RUFDATA paint a picture of extensive evaluation across the various cohorts of SEARCH participants. Aims and objectives were generally clearly set out and a variety of methods was used to collect data from all stakeholders. The evaluations demonstrate an evolutionary perspective with an early focus on programme development and as appropriate over time a shift in emphasis to programme impact. The programme was fortunate to have the resources to allow for the contracting of external evaluation experts who became familiar with the programme and with the faculty. As noted above there is a question, given their extensive contact with the programme whether these evaluators could really be considered as truly independent.

Table 15 Summary of RUFDATA details

Name	Method/Source	Aim/Objective (Reason/Purpose)	Uses	Foci	Data	Audience	Timing	Agency
SEARCH I	Phone Interviews @ 4, 6, 12, 18 and 24 months; email questionnaires SEARCH I participants, managers, selected RHA chairs and CEOs, selected faculty, advisory group members and AFHMR	To assess success of the program in meeting its defined objectives including: the use of evidence-based decision making in community health programming; the level of awareness and recognition garnered by the program at local, national and international level; and the satisfaction of the stakeholder with the program	To inform ongoing evaluation and modification of the SEARCH Program	SEARCH Program impact as reported by all stakeholders	Qualitative interviews; quantitative surveys	All stakeholders	Evaluations done throughout the first two years of the program	External: Barrington
SEARCH II	On line survey (6months); 6 training module evaluations; participant focus groups (12m); participant survey and FG(24m); supervisor phone interviews (24m) SEARCH II participants and direct supervisors	Aims and objectives varied for different parts of the evaluation depending on focus: e.g. participant feedback, supervisor feedback. Aims outlined for each evaluation activity. All data aimed to inform programme development for SEARCH III	To inform the improvement/ development of SEARCH III	SEARCH Program impact as reported by all stakeholders	Qualitative - interviews; quantitative surveys	All stakeholders	Throughout and up to 2 years following the programme	External: Health Informatics

Name	Method/Source	Aim/Objective (Reason/Purpose)	Uses	Foci	Data	Audience	Timing	Agency
SEARCH I and II Long Term Done by same company as SEARCH II evaluation	Survey document review literature review logic model development SEARCH I and II participant exploratory interviews and survey. SEARCH I and SEARCH II Individual participants	Assessing impact at individual level	Not stated	Individual participant impact	Survey responses	SEARCH administration	Following completion of SEARCH I (5 years) and SEARCH II (2 years)	External: McCaffrey Consulting
SEARCH III	Focus groups On-line survey SEARCH III participants	To formally assess the immediate and long term impact of the SEARCH program on individual SEARCH III participants including the application of skills in practice, the use of the SEARCH Network, personal and professional development and dissemination and application of findings emerging from SEARCH projects. To solicit feedback from participants regarding course content and delivery. To develop and refine processes for the ongoing evaluation of the impact of SEARCH at the individual participant level.	To assess individual impact and examine the role of the SEARCH projects.	SEARCH participants	On-line survey results	SEARCH administration and faculty	24 months post completion of SEARCH III	External: McCaffrey Consulting

Name	Method/Source	Aim/Objective (Reason/Purpose)	Uses	Foci	Data	Audience	Timing	Agency
SEARCH IV 12 months	Focus Groups SEARCH IV participants	The purpose of the project was to gather qualitative feedback from current SEARCH participants to help staff and faculty plan the remainder of the program, and to incorporate any changes into the final module and wrap-up if appropriate. Project findings may also inform the development of subsequent program iterations.	To inform current and future program development.	Participant views on: perceived impact and application of skills; individual and group projects (including progress, faculty support, and process for completion); value and relevance of curriculum design; and SEARCH awards and recognition.	Focus group interviews	SEARCH managers and faculty	Mid-term SEARCH IV	External: McCaffrey Consulting
SEARCH IV Long Term	On-line survey SEARCH IV participants	Same as for SEARCH III follow-up Note: Recommendations from this survey include consideration of data collected by the same consultants in the previous surveys of SEARCH II and II	Same as SEARCH III follow-up	SEARCH participants	On-line survey results	SEARCH administration and faculty	15 months post completion of SEACH IV	External McCaffrey Consulting

Name	Method/Source	Aim/Objective (Reason/Purpose)	Uses	Foci	Data	Audience	Timing	Agency
Faculty impact and experience	Document review, interviews, group interviews Core faculty team members	The purpose of this report is to capture the key dimensions of faculty engagement which are important to the short and longer term evolution and evaluation of the SEARCH program, and to consider the implications of those dimensions for program development and impact assessment	Program development and impact assessment	Faculty perspective	Individual and group interviews, document review	SEARCH administration and faculty	Post completion of SEARCH III	External: On Manage-ment Ltd
Project tracking	Primary data: In-person/telephone interviews and/or email. Secondary data: review of existing data sources, websites, and selected SEARCH project reports SEARCH participant managers/super-visors	To assess the extent to which SEARCH I and II projects have been applied or used in practice. To assess the extent to which SEARCH projects have made a difference for participants, their organizations, and the overall health system. To recommend measures and processes for the periodic assessment of SEARCH project findings.	Not stated	Mangers/sup-ervisors	Interviews, documents, project reports	SEARCH administration, faculty, participating organisations	End of SEARCH II	External: McCaffrey Consulting

Name	Method/Source	Aim/Objective (Reason/Purpose)	Uses	Foci	Data	Audience	Timing	Agency
Organizational Impact: evaluation 1 Combined project that included workshop	Interview Survey Workshops: All Alberta RHAs: managers	To determine to what extent involvement in the SEARCH Program resulted in some measure of change or outcomes for participating organizations. To what extent did SEARCH meet related goals of participating organizations? To identify organisation research capacity, develop evaluation conceptual framework and measure impact of SEARCH on participating organisations	To assess impact and inform program development	Participant/organization impact	Qualitative data from structured telephone interviews with individuals in participating organizations	AHFMR, SEARCH faculty, participating organizations	Post completion of SEARCH II	External: On Manage-ment Ltd
Mangers' Survey Only summary documents available	On-line survey Managers of SEARCH III, IV and V participants Based on Project Tracking Report 2003	To assess organizational impact of SEARCH Program	Unclear	Supervisors and managers	Quantitative survey based on 2003 survey	Unclear	During SEARCH V	Internal: Biddle
Collaborative Network Evaluation	On-line survey: Wilder Collaboration Factor Inventory SEARCH I - IV participants	To assess the degree of collaboration in the SEARCH Network	To inform Steering committee	Participants	On-line survey results of Wilder Collaboration Factor inventory	SEARCH Steering committee	End of SEARCH V	Internal: Biddle

5.4 Impact level results

A somewhat different lens with which to compare the evaluations is to examine the level at which the impact of the programme was being evaluated. To this end data were extracted from the evaluations in relation to the levels of impact outlined earlier from Saunders (2007) and presented in detail in the Methods section.

The data for these levels are presented in Table 16 and for ease of reference the aims and objectives of the evaluations have been repeated in this table. It is worth noting that the final evaluation listed in the table refers to an on-line collaboration inventory tool and so did not lend itself to data extraction in these categories and therefore the discussion refers to only ten evaluations.

Level 1 impact relates to the quality of the experience of the programme by participants. Of the 10 evaluations included in the data extraction, only three did not report Level 1 impact. Given the focus of these three evaluations this omission was considered appropriate. Both the participant and faculty focused evaluations reported positive personal experiences.

Level 2 impact relates to the quality of the situated learning outcomes, and of skills and knowledge acquisition. All ten evaluations presented data related to this level. Evaluations that focused on participants reported a consistent increase in knowledge and skills as a result of participating in the programme. In the faculty-focused evaluation there were reports of professional development that was attributed directly to faculty contact with the programme.

Level 3 impact deals with transfer or reconstruction of learning into the work environment. Changes in the roles and responsibility of the participants as a result of participation in the SEARCH Program were consistently reported across the evaluations. SEARCH participants took on leadership roles in relation to their clinical practice as well as in the area of conducting and using research findings in their respective institutions.

There were also reports of new research networks developing that included SEARCH participants and faculty as well as others within the health regions. These activities demonstrated an expansion of individual participant roles and an increase

in their confidence to explore the use of their newly acquired knowledge in a broader environment.

The faculty evaluation had a different focus and raised concerns regarding their academic roles within their home institutions (faculty were based in three different academic institutions). There were issues related to how their participation in the SEARCH Program was perceived by their supervisors and colleagues and the worth and merit of that participation. On a personal level there were also tensions with colleagues who, like the colleagues of the participants, felt that the faculty were just taking a week away while others stayed behind and carried increased workloads. Professionally, although the faculty stated that being involved in the SEARCH Program was important and worthwhile, there was also a sense that it did not directly contribute to the areas valued within the academic institutions (e.g. acquisition of grants and writing peer reviewed publications). This issue of academic merit was not resolved during the life of the SEARCH Program.

Organisational impact, the fourth level, proved to be much harder to measure. Early evaluations provided mixed reports of impact, and the conclusion was that it was too early to tell. Later participant evaluations provided examples of the influence of SEARCH participants on research use and conduct, development of collaborative networks (within and outside their home institutions). There were also less positive reports that reflected the disappointment that results of individual projects were not implemented, as well as the limited number of groups projects that were completed.

As noted earlier, within academic institutions, the participation of academic staff in the SEARCH Program was at times seen by some as a drain on resources. This was an issue even though the faculty time was purchased from the academic institutions. There appeared to be at least three points of contention. The first related to inconsistencies as to where funds were allocated internally and whether they were actually used to buy in replacement staff. In relation to the latter point there was the problem of actually being able to backfill the positions of the faculty participating in the SEARCH Program. The faculty were in senior positions and it would have been difficult to simply bring in new staff to cover workloads. There was therefore a sense amongst faculty that at certain times they were managing a full workload in their

institutions and doing SEARCH work, although there was variability across the faculty members. The other issue was the perception that faculty participation in the SEARCH Program actually took more time than was purchased by the programme.

Measuring any long term impact in the health care system was difficult, and linking this causally to the SEARCH Program was not possible in spite of the fact that long term evaluations were carried out. The project tracking evaluation that was carried out included only SEARCH I and II, and therefore was probably too early in the process to be able to measure impact on the participating organisations.

There were measures of publication of research and project findings. However, there was also concern regarding the limited dissemination of project findings and their limited impact. In terms of faculty, as noted above there was a sense that participation in the programme limited career advancement, owing to a decrease in grant income and peer reviewed publications. One of the key long term objectives of the SEARCH Program was the development of a collaborative network. None of the identified evaluations examined the existence or potential impact of such a network.

5.4.1 Value in using Impact framework

Use of the Impact framework was somewhat more problematic than that experienced with the RUFDATA framework. In terms of the evaluations that focused on the programme overall or on the faculty, data extraction was quite straightforward and category consistency prevailed. As should have been expected, not all evaluations reported impact at all levels, and it was not appropriate to attempt to make them fit all the categories. As noted earlier the Collaborative Network Evaluation did not fit any of the level categories. The project tracking evaluation reported positive experiences at the individual levels, but was focused more on the impact on the health authorities. The focus of the Organizational Impact Evaluation meant that by definition impact on individuals was not reported.

As a group, the evaluations covered the entire range of impact levels but as would be expected the impact at the organisational level, whether health authority or province, would take time to accrue. Consequently, establishing a causal link or even correlation in the complex and changing environment of health care delivery proved to be difficult.

Table 16 Summary of Impact level data details

Name	Aim/Objective (Reason/Purpose)	Level 1	Level 2	Level 3	Level 4	Level 5
SEARCH I	To assess success of the program in meeting its defined objectives including: the use of evidence-based decision making in community health programming; the level of awareness and recognition garnered by the program at local, national and international level; and the satisfaction of the stakeholder with the program	Participants rated the experience positively although they found the 7 week course intense and difficult to organise with work and home commitments. Reports of ongoing support were mixed.	General feeling that the learning was valuable to participants and allowed them to expand in their professional roles	Participants and supervisors felt there had been a transfer of awareness of the need for the use of research findings in local contexts and the need for good quality local research.	Mixed reports of the impact of the SEARCHER on institutional setting although the sense was that it was too soon to tell.	No long term impacts reported
SEARCH II	Aims and objectives varied for different parts of the evaluation depending on focus: e.g. Participant feedback, supervisor feedback. Aims outlined for each evaluation activity. All data aimed to inform programme development for SEARCH III	Participants rated the experience positively and recommended implementation changes	Participants identified role changes due to improvement of their research knowledge	Participants reported using new skills as part of their professional roles. Supervisors reported integration of new knowledge by participants.	Concrete examples of influence on research and EBP activities in the workplace	No long term impacts reported
SEARCH I and II Long Term Done by same company as SEARCH II evaluation	Assessing impact at individual level	Reported positive experience	Increased individual research skills, professional networks, professional advancement	Use of research and leadership skills in work environment	Increased skill of workforce, development of collaborative networks, facilitation of research, improved information retrieval, policy changes due to projects	Participation in strategic planning, research publication - 64 external and 15 peer-reviewed journal publications

Name	Aim/Objective (Reason/Purpose)	Level 1	Level 2	Level 3	Level 4	Level 5
SEARCH III	To formally assess the immediate and long term impact of the SEARCH program on individual SEARCH III participants including the application of skills in practice, the utilisation of the SEARCH Network, personal and professional development and dissemination and application of finding emerging from SEARCH projects. To solicit feedback from participants regarding course content and delivery. To develop and refine processes for the ongoing evaluation of the impact of SEARCH at the individual participant level.	SEARCH rated as a positive experience by participants	Reported increase in knowledge base - especially searching skills, research knowledge, use of networks. Reported increased use of on-line learning facilities	Reported the use of new skills in work environment including taking increase responsibility and leadership roles in decision making. Reported continued networking within and outside the SEARCH networks	Reported some disappointment in the lack of wide dissemination of project results.	None reported
SEARCH IV 12 months	The purpose of the project was to gather qualitative feedback from current SEARCH participants to help staff and faculty plan the remainder of the program, and to incorporate any changes into the final module and wrap-up if appropriate. Project findings may also inform the development of subsequent program iterations.	SEARCH continued to be rated as a positive experience by participants	Similar findings to SEARCH III	Reported continued use of skills in practice. Terms changed slightly with introduction of 'scholar practitioner' and 'change agent'	Individual projects used internally only. Limited number of group projects complete.	None reported

Name	Aim/Objective (Reason/Purpose)	Level 1	Level 2	Level 3	Level 4	Level 5
SEARCH IV Long Term	Same as for SEARCH III follow-up	Reported positive experience and support for current curriculum	Reported acquisition of new knowledge and skills that were appropriate to their work	Reported using their skills in their work environment although this was not the focus of this evaluation	None reported - not the focus of this evaluation	Commented on desire to have project findings more widely recognized and disseminated
Faculty impact and experience	The purpose of this report is to capture the key dimensions of faculty engagement which are important to the short and longer term evolution and evaluation of the SEARCH program, and to consider the implications of those dimensions for program development and impact assessment	Faculty report satisfaction in their role in SEARCH	Report important links with health authorities and other researchers	Report a lack of recognition of their work within their academic departments	Report that SEARCH work is seen as a resource drain to their academic departments and took more time than anticipated	Report their SEARCH work as contributing to their personal and professional goals but not contributing to overall academic roles.
Project tracking	To assess the extent to which SEARCH I and II projects have been applied or used in practice. To assess the extent to which SEARCH projects have made a difference for participants, their organizations, and the overall health system. To recommend measures and processes for the periodic assessment of SEARCH project findings.	Participants reported experience of projects and additional projects not previously listed	Completion of projects and implementation demonstrated extent of participant knowledge base	Completion of projects and implementation demonstrated impact in work place	Reports of implementation of projects at local level	Limited impact on decision making for strategic objectives

Name	Aim/Objective (Reason/Purpose)	Level 1	Level 2	Level 3	Level 4	Level 5
Organizational Impact: evaluation 1 Combined project that included workshops	To determine: To what extent involvement in the SEARCH Program resulted in some measure of change or outcomes for participating organizations? To what extent did SEARCH meet related goals of participating organizations? Factors that are predictive or suggestive of success?	Not reported	Perceived positive impact on participants	Increased knowledge of SEARCHERs demonstrated in their roles and activities	Perceived organizational impact on priority setting, collaboration, networking, identification of information and research activities	Limited impact on decision making for strategic objectives
Mangers' Survey Only summary document available	To assess organizational impact of SEARCH Program	Not reported	Reported increased skill level of participants	Reported use of skills in workplace by SEARCH participants	Reported use of skills to inform decision making in the workplace	Reported differences in definition of dissemination of results of SEARCH projects
Collaborative Network Evaluation	To assess the degree of collaboration in the SEARCH Network	N/A	N/A	N/A	N/A	N/A

5.5 Quantitative data

As noted in the Methods section each evaluation was examined, and judgements were made in relation to the checklist provided by Stufflebeam (1999). The initial data extraction form was pilot tested using three evaluations. It was ifound that the cohort evaluations individually did not include all the information regarding the conduct of the evaluation process, and a decision was therefore taken to group the cohort evaluations together (as described earlier) for the purpose of quantitative analysis.

Data were extracted by one reviewer on two separate occasions into two separate data extraction forms for three of the included evaluations. If no mention was made of an item then it was scored as 0. Results of the two sets of data extractionwere compared and where the assignment of the scores did not match the evaluation was re-examined and a final decision made. On average there were 12 discrepancies over the 300 data points (range 10-14) in each evaluation, demonstrating correlation between the assessments. Therefore this data cross checking process was not repeated for the remainder of the evaluations. The individual evaluation scores were so low that even if this level of reproducibility was repeated across all the evaluations it would have had a very limited impact on the overall scores or the conclusions drawn.

Scores across all standards were so low that a descriptive analysis proved of limited value. Therefore this information has been placed in Appendix 4 which includes a narrative description of the assessment results for each standard and a table of the summative results. An overview is provided below. Copies of the detailed data extraction tables are available on request.

As a reminder for the reader there are totals of 7, 3, 8 and 12 items respectively in the four standards of utility, feasibility, propriety and accuracy and each item can have a maximum score of ten. The strength of the evaluation was determined using the formulae provided within Stufflebeam's metaevaluation tool (see Appendix 2). The metaevaluation tool uses formulae to determine the strength of the evaluation based on the proportion of excellent to poor ratings and converts this to a proportion

out of 100. Evaluations were also appraised according to his pre-specified pass/fail criteria.

5.5.1 Quantitative analysis

Utility data

The utility standard relates to the ability of the evaluation to meet the information needs of the intended users, and is made up of seven checklist items.

The only item on which the evaluations rated well was the fifth, which measured report clarity. Two reports scored of 6/10 with the remainder scoring 9 or 10. The reports were professional in presentation, well organised and clearly written.

The high rating for report clarity item meant that a number of evaluations scored at least one excellent mark in their overall score. However, in general the ratings were only fair thus producing strength scores that ranged from 3 to 12 and an overall strength score that ranged between 11% and 43%. The Managers' Survey and the Collaborative Network Survey had the lowest scores in this category – a situation that is repeated in all the other standards.

Feasibility data

This is the shortest of the four standards with only three items, and relates to whether the evaluation is realistic, prudent, diplomatic, and frugal.

In terms of feasibility there were no excellent or very good scores, and the strength scores were between 2 and 3. In terms of overall results the scores ranged from 16.7% to 25% indicating a very low strength of the evaluations.

Propriety data

There are eight items in the Propriety standard, which assess the legal and ethical issues related to the evaluation by examining whether there was due regard for those involved in the evaluation or affected by its results.

Overall in this category there were no scores of excellent or very good, and the overall percentages ranged from 3% to 22% indicating a very low strength of the evaluations' provision for propriety.

Accuracy data

The twelve items in this standard relate to technical adequacy of the report in relation to the programme under evaluation including a determination of the merit and worth of the programme.

Overall only three reports included an item that received an excellent rating. For the remainder, the scores were predominantly poor. The accuracy strength rating ranged from 4 to 12 with the majority (6) scoring over 10. However, this resulted in consistently low strength scores that ranged between 8% and 25%.

5.5.2 Failure categories

Stufflebeam recommended that a score of poor (0-2) on any of four specific items (P1, A5, 10 and 11) from the standards criteria should mean that the evaluation failed. As can be seen in Table 17 all the evaluations scored poor in A11 (Impartial reporting) and therefore would be considered failures. A number of the evaluations failed in more than one item and the Managers' survey failed on all of the designated critical criteria.

Therefore overall the key evaluations conducted as part of the development and delivery of this programme scored poorly using the criteria set by the Joint Committee on Standards(1994). In fact using the pass/fail criteria none of the evaluations reached that critical level. Further discussion of the validity of the tool and the reasons for these low scores is presented in the next chapter.

Table 17 Failure Categories

Reference	P1 Service Orientation*	A5 Valid Information	A10 Justified Conclusions	A11 Impartial reporting
SEARCH I	4	3	3	0
SEARCH II	4	4	4	2
SEARCH I and II	3	6	1	0
SEARCH III	4	6	1	1
SEARCH IV	3	6	4	0
Faculty Impact	4	2	3	1
Project Tracking	2	5	3	1
Organisational Impact 1	3	1	2	1
Managers' Survey	1	2	1	0
Collaborative Network Evaluation	0	3	1	0

*Scores out of 10 for ten evaluation tools, in the four categories listed in column headings, based on criteria set by the Joint Committee on Standards (1994)

5.6 Developmental evaluation analysis

This section discusses the available evidence in an attempt to demonstrate that the SEARCH Program was a complex programme functioning in a complex environment. It then goes on to explore whether it meets the criteria set out by Patton (2011) for the use of developmental evaluation. It goes on to present data to demonstrate that although the evaluation activities were not specifically defined as developmental evaluation by the faculty and evaluators, the conduct and results of the evaluations played a key role in programme development.

5.6.1 Complexity and the SEARCH Program

Prior to making a decision regarding the appropriateness of using the developmental evaluation lens to examine the evaluation processes used within the SEARCH Program it is necessary to determine whether the SEARCH Program fits the definition of a complex, evolving and innovative programme. Analysis of this has been done from two perspectives, Patton's (2011) complexity concepts and the Stacey matrix (2002).

Table 18 presents Patton's concepts of complexity and information regarding the SEARCH Program, demonstrating it meets Patton's descriptions of complexity.

Table 18 The EBP and the SEARCH Program as complex environments

Complexity concepts	**Description***	**SEARCH Program**
Nonlinearity	Sensitivity to initial conditions; small changes have major impact (e.g. movement of butterfly wings)	The SEARCH Program was sensitive to concerns about whether health authorities would accept staff capacity development as necessary to move forward the EBP agenda. It was clear that some health authorities were moving more quickly than others and it was not possible to predict even within them which clinical areas would see the need for staff development. A change in local leadership was seen to affect the acceptance of the programme concepts both positively and negatively..
Emergence	Patterns emerging from self-organisation among interacting agents	The SEARCH Program was emergent – although based on the concepts of INCLEN it none the less needed to evolve its own curriculum, faculty and method of delivery. This required the establishment of new relationships with both the health authorities and the universities from which faculty were recruited.
Dynamical	Interactions between and among subsystems which may be volatile, turbulent, cascading rapidly and unpredictable	The evolution of the SEARCH Program itself was dynamic and changing on a number of fronts (curriculum, programme delivery, faculty etc).
Adaptive	Interacting elements and agents respond and adapt to each other	The movement of SEARCHERs in and out of the SEARCH Program and their health authorities required constant adaptation on the part of organisations, participants, faculty and the SEARCH Program.
Uncertainty	Processes and outcomes are unpredictable	There was constant uncertainty as to the reactions of those in the health authorities to outcomes of both the implementation of EBP and the SEARCH Program The changing delivery of the SEARCH curriculum also meant that responses of participants remained uncertain. As time moved on there was also uncertainty regarding funding of the programme.
Co-evolutionary	Interactive and adaptive agents evolve together within and as part of the whole system	Both the health care system and the SEARCH Program systems were evolving individually and together.

*Adapted from Patton (2011) pg 8

However this is a single perspective and there are other lens that would help to either confirm or deny that this programme and environment should be considered as complex. Comparison of the attributes of the SEARCH Program within the Stacy

complexity matrix have been uset to examine whether it was functioning in the 'zone of complexity' (Stacey, 2002).

Stacey matrix

The Stacey matrix emerged from the management literature and was designed as a management decision tool (Stacey, 2002). iIt hinges on two aspects of decision making; certainty and agreement (see Figure 3)

The horizontal axis shows how certain we are about the chances that the current course of action is the correct one – that is we have experiential knowledge that the current plan of action will cause an anticipated result. In the case of the SEARCH Program there was no certainty that the proposed programme, would meet the established goals or the requirements of the health authorities. There was evidence from the success of the INCLEN (International Clinical Epidemiology Network) programme that such a teaching/mentoring model had worked in the field of international clinical epidemiology – but EBP was a much more uncertain area and therefore the results could not be predicted.

The vertical axis deals with the agreement across all those involved about the desired outcomes. When the SEARCH Program was first proposed as one approach to address the challenges of introducing EBP there was agreement that providing the best patient care was paramount. However, there was very limited agreement about how that would be accomplished or measured. In fact the various members of the health care community were only just beginning to come together to discuss the issues. Therefore in terms of the Stacey complexity matrix the SEARCH Program demonstrated both uncertainty and a lack of agreement of process and outcomes and so would be considered to be operating in the 'zone of complexity' or, in terms of the diagram, in the area requiring 'complex decision making'.

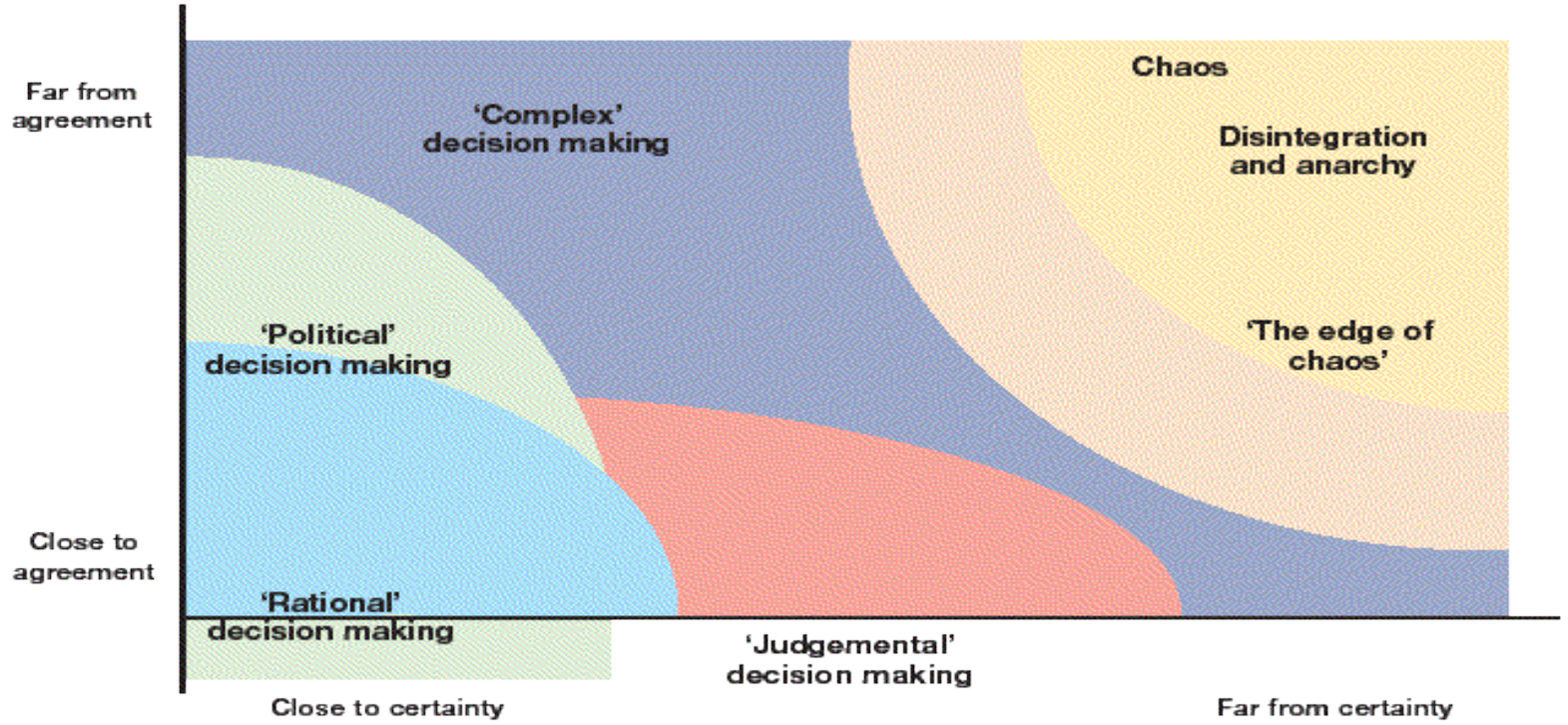
Although all this was true for the programme overall, examination from a different perspective, that of the then director of AHFMR - might point towards a somewhat different conclusion (personal communication M. Spence, 1996). Dr. Matt Spence stated in a number of public forums that he was confident that an active CPD programme was one of a number of possible approaches to moving forward the implementation of EBP in Alberta. He was also convinced that this would require an interdisciplinary educational programme and the development and promotion of

networks of health care workers across the province. He was certain that it could be done because he had control of the funds to make it happen. So because of his position and confidence he could mandate agreement on the outcomes. However, that did not mean that he moved forward without consultation. His consultations were wide (international, national and local) and it was his nature to work in active collaborations. His approach was nevertheless what has been referred to historically in management literature (Peters and Waterman, 1982) and later referred by Patton (2011) as 'ready, fire, aim'.

5.6.2 Developmental Evaluation and the SEARCH Program

This is all evidence that the SEARCH Program was a complex programme being established in a complex environment. However, it does not demonstrate that the use of developmental evaluation would have been appropriate. The literature review presented earlier included the questions posed by Gamble (2008) to determine whether an environment was appropriate for the use of developmental evaluation. These are presented again in Table 19 and include the perspective of the SEARCH Program.

Figure 3 Stacey complexity matrix



Adapted from Stacey (2002)

Table 19 Gamble's environment questions

Question and Rationale	**SEARCH Perspective**
1. What is driving the innovation? Developmental evaluation is particularly appropriate if an organization expects to develop and modify a program over the long term because of constantly shifting needs and/or contexts. It is helpful to differentiate between innovation taking place within an organization and the adoption of an external innovation, which may not need a developmental evaluation.	There was no question in the mind of the SEARCH Program initiators that the process would evolve over a period of time and that the environment in which EBP was being implemented was uncertain and changing. It could be argued that the innovation was an adaptation of the INCLEN programme and therefore it did not need developmental evaluation. However, although the INCLEN programme served as a model there were substantive differences in both the environment and the implementation that meant the SEARCH Program was evolving within the dynamic situation of EBP
2. Are the proposed changes and innovations aimed at deep and sustained change? Developmental evaluation is aimed at innovations that are driving towards transformational changes. Organizations often fine-tune their programs, and having an evaluative lens on those changes can be helpful; however the intensity of developmental evaluation may not be warranted in every instance.	The implementation of EBP was a transformational change in health care delivery and could be termed a paradigm shift. This was the environment in which the SEARCH Program was established. As a model for CPD delivery it was also innovative and underwent significant changes over time. As a programme SEARCH was also innovative and as such was subject to major transformational changes in the first five iterations of the programme.
3. Do we have a collaborative relationship with another organization in which there is innovative potential in combining our respective talents? Developmental evaluation may help different organizations work together through the effort to innovate. In this situation, the developmental evaluator can help the organizations through some of the inevitable tensions of collaborating and can provide a measure of transparency about the experiment.	Collaborations were developed with educational institutions, the regional health authorities as well as other health boards (eg. Mental health board) and institutions. The requirement of individual and group projects within the SEARCH Program required collaboration across all these groups.

Question and Rationale	SEARCH Perspective
4. Under what conditions does the organization currently innovate? Is innovation part of the culture of the organization? If this is already part of the culture, then the developmental evaluation role may be one that people within the team already play. If there is not a culture of innovation but there is a commitment to build one, then developmental evaluation may be helpful in stimulating that.	The SEARCH Program in and of itself was an innovation. It was accepted from the inception of the programme that on-going evaluation and innovation would be part of the SEARCH Program.
5. What are some core elements of what we do that we don't want to change? There may be elements of an initiative that are known to work, or for another reason are expected to stay the same. Evaluation requires resources, and if things will not change, these resources are better directed elsewhere. If something is not going to be adapted but there is interest in finding out whether it works, a summative evaluation is appropriate.	There was an open acceptance that all aspects of the programme were open to evaluation and change. Some more than others – e.g. the three themes of teaching were relatively unchanged over time but all other aspects of programme promotion and delivery changed over time.
6. Is it clear for whom the evaluation is intended? This is a vital question for any evaluation, developmental or otherwise. For an organization to make good use of developmental evaluation, it is important to have key decision makers interested in and open to using evaluative feedback to shape future actions. If the only user of the evaluation is external to the innovating team (such as a funder), then developmental evaluation is probably not the appropriate approach	The evaluations were done to inform the future development of the programme and to inform all stakeholder (participants, faculty, AHFMR and health authorities).

(Gamble, 2008)

Given this evidence, it is clearly demonstrated that the SEARCH Program was innovative, evolving and functioning in a complex environment. Comparison with Gamble's questions also demonstrates that it meets the criteria set for consideration of the use of developmental evaluation. However, developmental evaluation is carried out for a number of different reasons.

Patton (2011) outlines five possible purposes for developmental evaluation (Table 20). The SEARCH programme fits with the initial and final purposes on this list. It was an ongoing development that was adapting an innovative initiative within a complex environment. It also was part of two major system changes. Firstly it was designed to assist in the adoption of EBP, and secondly as a CPD approach it was also a step change away from what had been the norm.

Table 20 Purposes and uses of developmental evaluation*

Purpose	**Use**
Ongoing development	To adapt an innovative initiative to new conditions in complex dynamic systems
Adapting effective general principles	The use of ideas or innovations taken from elsewhere to be developed in a new setting
Developing a rapid response	In cases of major change or crisis to explore real-time solutions and innovations
Performative development of potentially scalable innovation	To bring innovative programs to the stage where they are ready for formative or summative evaluation
Major systems change and cross-scale developmental evaluation	Providing feedback regarding the evolution of major change and how this might impact on the broader dissemination of a project (horizontal and vertical scaling)

*Adapted from Patton (2011) pg 21-22

It could also be argued that it met the second purpose of adapting an innovation from somewhere else. However, the links with the INCLEN Programme were not maintained and the SEARCH Program adopted a very different approach as it evolved over time and therefore this has not be considered as one of its purposes.

5.7 Qualitative data extraction

As outlined in the methods section, qualitative data were extracted from SEARCH Program documents. These include minutes of the SEARCH Steering Committee and Evaluation Steering Committee meetings and accompanying documents for the period from 2000 to 2005. A qualitative directed content analysis approach was used (Krippendorff and Bock, 2009). The purpose of this approach was to

address three issues. The first was to provide supportive evidence that the SEARCH Program was functioning in a complex environment that required working collaboratively with a variety of stakeholders. The second was to demonstrate that the administrators, faculty and programme committees had an embedded evaluative culture and that the focus and use of the evaluations were consistent with what Patton (2011) refers to as developmental evaluation. The final consideration is the investigation of whether the changes that were made in the programme were dramatic enough to be considered within developmental evaluation as opposed to less significant changes that you would expect with standard formative evaluation.

Data were extracted in these three categories using eight codes. The categories are presented in Table 21 which also provides a definition and purpose for each code. The codes relating to the complex environment were linked to integration with the health care system, with AHFMR and the SEARCH Program faculty. Codes relating to evaluative culture included evaluation approaches, culture and use. The final and largest code related to innovations in the programme, and was used to identify evidence that the changes that were made were not minor, but represented significant alterations in the programme, thereby making the use of developmental evaluation appropriate. These are discussed in relation to programme delivery, faculty and external programme contributors.

5.8 Qualitative data analysis

In this section the data relating to the complexity of the environment will be presented in relation to the health care delivery system, to AHFMR and to the SEARCH faculty. This is followed by the analysis of the developmental evaluative ambience of the SEARCH Program, its evaluative practices and a description of the extensive changes made to the programme as a consequence of the evaluations. Data quotations are followed by the source document details.

Table 21 Qualitative coding categories

Code	Code definition	Purpose
Environment		
Integration – Health Care System	Any description of linkages with health authorities	Evidence of the complex and evolving world in which the programme and participants worked
Integration - AHFMR	Description of integration/relationship with AHFMR and AHFMR goals	Evidence of links/integration with host organisation
Relationship - Faculty	Any description of faculty roles and role changes within the programme or within their institutions	Evidence of evolving programme links with faculty
Developmental evaluation culture		
Evaluation Approach/Practice	Any description of the general approach taken to evaluation	Evidence of actual evaluation approaches. Evidence of a wide range of evaluative practices – triangulates with report findings
Evaluation Culture	Any comments related to the importance and role of evaluation	Evidence of the value placed on evaluation
Evaluation Use	Any description of the attitudes to evaluation outcome and use	Evidence of the use of evaluation in programme development and delivery
Programme innovations		
Innovation-Programme	References to changes made in programme curriculum, delivery , faculty etc.	Evidence of the evolution and changes made in the programme

5.8.1 Environment

Health care system

As previously discussed, introducing the use of best evidence into the delivery of health care services was not, and is not, a straightforward matter. In addition, the province of Alberta is geographically large and was at the time divided into a number of different health regions. The number of regions varied over time – with 17 regions when the SEARCH Program began, but was reduced down to nine regions during SEARCH III and ultimately to one during 2009. Each of these regions had different population needs and varying management approaches. The province is generally considered wealthy but there is a well-established north-south divide, with the focus in the north on the extraction of natural resources and the south on the management of those resources. Therefore although there was a provincial health minister that directed the decisions in the province there was significant variation across regions.

The relationship between the SEARCH Program and the health care system evolved over time. Initial contacts were made by the director of AHFMR and then later the SEARCH Program director. These contacts were at the level of the Chief Executive Officer (CEO). It was initially felt that a CEO from the health region would provide the supervision of the SEARCH participants. As a natural progression it was also assumed that at least one CEO would be on the SEARCH Steering Committee.

However as the programme evolved it was recognized that the CEO was too far removed from where the SEARCH participants worked and so their supervision should change as well as the role of the appointee on the Steering Committee;

> It was agreed that SEARCH participants do not have to be in a direct reporting relationship to their CEO. (SEARCH Steering Committee Minutes 0900)
>
> The committee discussed effective membership and endorsed the importance of senior Regional Health Authority (RHA) executive representation, while acknowledging that CEOs may not be the only appropriate participant. (FINAL MINUTES OF MEETING 082901)

This issue also affected how SEARCH participants were recruited. Initially it had been done at the level of the CEO. However as the quote below indicates this changed over time.

> Concern was expressed that, if organizations had to jump through too many hoops early in the process of joining SEARCH, this might act as a deterrent. A facilitated discussion with the organization, followed by a letter of understanding might work better. Issues should be discussed with the participant's supervisor or an appropriate liaison to the CEO as it will be the people in the organization closest to the participant who will need to provide the most support. The CEO, however, must be kept informed and be supportive of the process.(SEARCH Steering Committee Minutes 0900)

The quote also demonstrates the balance that was required to facilitate health authorities' participation and the need for ongoing communication with leaders in the health regions. This changed over time as the role of SEACH projects changed and decisions regarding projects were taken jointly by the health authority supervisors, the SEARCH participant and the SEARCH faculty mentors.

On the same note there was a constant struggle related to the organisational support and time allowed to SEARCH participants.

> SS introduced this discussion by summarizing the efforts made over the past three years to address the question of organizational support from health regions for SEARCH participants. There continues to be a tension identified by participants, in particular related to the time protection needed to focus on projects. (MINUTES OF MARCH 20 MEETING of Steering Committee)

Two other changes took place that provided positive links with the health regions. The first of these was initiated by the SEARCH Program when they included managers of SEARCH participants in an orientation meeting at the beginning of SEARCH III. As can be seen in the quote below this was repeated in SEARCH IV with positive results.

> SS reported on the successful SIV Managers' Orientation. She noted a real shift in the interest and perspective since SEARCH III. There was a shift in language with a focus on clarification of the manager's role. Messages that came across clearly at the March 19 Meeting were:
>
> - How do I support my participant?
> - What can we do to help?(DRAFT MINUTES OF MARCH 20 MEETING of Steering Committee.doc)

The second of these involved the invitation from a health region to hold a SEARCH Program module in their region.

> For the first time, the SEARCH program has been invited by a health authority to hold a module in their region. In June, Module VI will be held in Slave Lake (Keeweetinok Lakes RHA), (STEERING COMMITTEE March 7minutesfinal)

This change contributed to the overall goal of developing the SEARCH network and also allowed for opportunities to involve local CEOs, managers and even local politicians to become more familiar with and involved in the programme. From this point on all modules were held in different heath regions across the province.

As noted earlier the SEARCH Program was conducted in a province with a health care system that underwent two major re-organisations during the life of the programme. Data from only one of these changes are used in this evaluation, as the second occurred at the time the SEARCH Program ended. The restructuring events are reflected in the steering committee minutes and indeed they were felt to be so important that restructuring was given a standing position on the agenda under the heading of 'Environmental Scan'. Within this agenda item, members of the committee reported updates on the changes that were occurring within their newly defined regions.

I can report that I attended two modules with the SEARCH III cohort. The second module occurred at the time of the re-structuring, and of the 24 participants, four had been made redundant and a further six were facing possible redundancy. As noted earlier, evaluating the impact of the SEARCH Program on the development of networks across the province, was difficult. However, during this module such networking was very apparent. Session schedules were adjusted and time was provided for participants to discuss their current situation with support and alternative plans of action provided by other participants. Possible solutions were discussed and names of possible contacts external to the SEARCH Program were provided as possible leads to new employment opportunities for the affected participants.

It is clear that the SEARCH Program was working within a complex and evolving health care system. This required SEARCH Program leaders to establish close relationships with leaders in the health regions to ensure the most positive learning experience for participants as they continued their studies and carried out their project work. The invitation to hold SEARCH modules within the various health regions and the local support that this required is indicative of the acceptance of the goals and aspirations of the programme and the role that could be played locally to achieve those goals.

AHFMR

As noted earlier it was the vision of Dr. Matt Spence, AHFMR director, that launched and supported the SEARCH Program. It was his international experience and innovative thinking that brought the programme into existence. It was also his stalwart support within AHFMR that provided the continued vision of integration within the health care system and also the not insignificant financial support that the programme required. The cost for the first SEARCH cohort in 1996 was estimated at $1 million (CAD). This translated into approximately $40,000/participant (personal communication M. Spence, 1998). These were the direct costs to AHFMR and did not include the contribution of the health authorities relating to participant time and support for individual and group projects.

However, the SEARCH Program was only one of many AHFMR health research activities in the province. There was therefore always the need to ensure that there was an alignment of the aims of the SEARCH Program with the broader aims of AHFMR.

Examination of the Steering Committee meeting minutes identified a tension around the committee's role and authority. The extent of this uncertainty was also demonstrated by the fact that it took over a year of bi-monthly meetings for their terms of references to be accepted by the Steering Committee and sent back to the AHFMR board for approval.

> The committee indicated that there was a synergistic relationship between the Program and Health Authorities, and that the AHFMR Trustees have the ultimate say in what is accepted and that the role of the Committee is one of an Advisory one. (SEARCH Steering Committee Minutes 0700)

The mission of AHFMR was to improve health within the province through the conduct of high quality health care research. On the positive side this provided the impetus for the constant evaluation of the SEARCH Program. However, the other side of the coin was that research conducted within the SEARCH Program was not directly funded, nor overall was it at the same level of sophistication or impact as the majority of the medical research being funded by AHFMR. An extensive meeting discussion took place that examined the role of project funding and peer review (both aspects included in AHFMR funded research projects).

> AA reiterated AHFMR's mandate and goals from the AHFMR Act, 1980, and the June 1992 Strategic Planning Report, respectively; and the Foundation's commitment to maintaining Alberta's lead in health research. (SEARCH Steering Committee Minutes 0700)

> In conclusion, there is significant doubt about the overall value of any additional funding for SEARCH projects, and specifically about implications of a peer-review process. However the committee felt that there could be important value in looking at the question of improved quality and additional funds in a larger context, connected with other areas of programming - such as the SEARCH network, or the possibility of some support for SEARCH participants' time. The committee agreed that SS should take this discussion back to the Foundation and consider the issues and options broadly. (FINAL MINUTES OF MEETING 082901)

In the end there were attempts made to improve the level and impact of the participant research projects and a mechanism for seed funding for projects was established. However, there was a constant tension related to the purpose of the projects. That is, were they learning tools for the students or were they research projects designed to inform the development of health authority policy, or could they be both? As an example a project that was not completed or did not result in the hoped for results might be considered as a failure in relation to effect on health policy. However, the experience gained by the student during such a project could be very valuable and the knowledge gained could then be used in the development and management of future projects. The tension caused by these two alternate objectives was never resolved.

It is clear that the relationship between AHFMR and the SEARCH Program was complex, with funding, overall direction and support coming from the AHFMR and day to day operations management left to the SEARCH Program director and the various advisory and curriculum committees.

SEARCH faculty

The relationship between the programme and the faculty was no less complex. In the first instance the role and structure of the faculty changed significantly through the various iterations of the programme. The majority of the faculty for the first two cohorts were external to the programme and even the province (this was a reflection of the use of support from the INCLEN Programme). However, it became clear from the evaluations that a more coherent approach was required and therefore an investment was made to establish a core faculty. This required collaboration across a variety of academic disciplines (e.g. health services, health

economics, statistics, business and nursing) in the two largest universities in the province as well as more limited collaboration with two of the smaller universities. In addition a partnership was established with a private consultancy service and an individual from that organisation became a member of the core faculty for the duration of the programme. There was also a complex relationship with the institute within one of the universities that provided the development and support for the use of computer and informatics technology in the programme – this is discussed in more detail later in this chapter. So there evolved a complex situation of a core faculty from a variety of different disciplines and institutions across the province.

In general the faculty members were very positive about their work within the programme as demonstrated by this quote from Evaluation Steering Committee meeting.

> It became clear from the interviews that individual faculty got great personal and career satisfaction from contributing to the training and education of health professionals in the community setting, and in promoting the effective use of evidence and research to improve decision-making in organizations. (MINUTES - JANUARY 8 EVAL STEERING COMMITTEE)

However, there were indications that their roles in the SEARCH Program were different from those in their university environments.

> There was a discussion concerning the essence of the faculty role and how it differs from the role of a supervisor in any graduate program. We encourage the participants not to think of themselves as students, and that the SEARCH program is not based on any hierarchical learning community. Feedback from the participants has indicated (favorably) that the relationship they have with SEARCH Faculty is unlike relationships that they've had with faculty in the past. (MINUTES - JANUARY 12 2004)

Although this was seen as positive for the programme it meant that faculty were moving in and out of teaching environments that had very different philosophies and approaches.

The core faculty took active roles in the development of the curriculum through Faculty Committees. Although each curriculum theme (Creating, Choosing, Using) was chaired by a different faculty lead member, there were indications that the faculty themes were being integrated in practice.

> BB reported that, in the past, there's been much discussion at these meetings(Faculty Committee) regarding integration across curriculum, themes, The last module (Module 3) was an example of seeing that it's working - Theme integration was achieved through:
>
> - The use of a unifying case (for example: childhood asthma)
> - Faculty members teaching across themes
> - SEARCH peers providing teaching sessions (past participants from Chinook and DTHR which increases integration across SEARCH cohorts)
> - Ongoing attempts to connect with the organizations in the areas where we are holding the module. (MINUTES - SEPTEMBER 18 2003)

In addition to this the faculty themselves were in a unique position. Their university departments valued the link with AHFMR.

> The value for Departments/Faculties appeared to be in the good will of relationships with AHFMR, and in the funding received. (MINUTES - JANUARY 8 EVAL STEERING COMMITTEE)

However the data reflect some of the issues raised in the formal evaluation of the faculty. That is they were working in an exciting and innovative programme, which they enjoyed and they felt was contributing in a substantive way to the continuing professional development of health care professionals. However, the outputs from this work did not contribute to their academic responsibilities related to conduct and publish the results of high quality research, nor was it, in some instances, counted as contributing to their individual teaching load in their home institutions even though funding arrangements were in place.

> The biggest issues for faculty, emanating from the interviews, was the issue of "traditional" performance measures relating to career progression and performance in the academic setting, and direct control over the funding provided. None of the faculty interviewed had any research or publications (peer reviewed or otherwise) resulting from their involvement in SEARCH. (MINUTES - JANUARY 8 EVAL STEERING COMMITTEE)

There were many more examples of the complexity of the environment in which the SEARCH Program functioned. However, the data already presented substantiate the claim as well as demonstrating that there was a constant challenge to balance the needs of the participants, the health regions and the faculty.

5.8.2 Evaluative practice

As seen in Table 21 three codes were used in to identify and demonstrate that the conduct of evaluation and response to the results of such evaluations was embedded in the culture of the SEARCH Program. The existence of three different codes for evaluation approaches, evaluation culture and evaluation use

did not prove to be particularly useful as there was significant overlap between categories.

In terms of approaches to evaluations, it has already been shown from the quantitative data that a variety of evaluation designs were used in the programme evaluations. The excerpts provided below address the way in which results of evaluation were embedded in all discussions and used for future programme planning.

It is interesting to note that there were very few minutes from the Evaluation Steering Committee as it seems that following the setting of the initial direction for evaluation it was less active. There are two possible reasons for this. The first is that there is evidence that they made detailed recommendations to the Steering Committee regarding the development of an 'Evaluation Blueprint' which directed the evaluation activities of the SEARCH Program over a ten year period. Therefore there was less reason for them to meet if the evaluation plan was evolving as it was meant to. In addition, members of the Evaluation Steering Committee also sat on the overall programme Steering Committee and this meant that evaluation issues could be managed at that level.

> On the advice of the Evaluation Steering Committee, CC and DD were engaged to develop an 'Evaluation Blueprint", to inform the coordination and synthesis of information about the SEARCH program and its impacts over the next ten years. They were to describe the scope for future and past evaluation activities, capture the conceptual models developed by the committee and develop a road map to identify priority actions for the future.
>
> CC presented an overview of the Blueprint and highlighted that the evaluation and program design processes are intertwined. Therefore the scope of the document needed to address program design as well. The Blueprint is a means of stepping back to say, "where are we now and where to from here"?(FINAL MINUTES OF MEETING 082901)

It is important to note however that the Evaluation Steering Committee set the approach to be taken in relation to assessing impact on sponsoring organisations.

> The general approach includes developing a generic framework relating to organizational capacity for doing and using research and then using this framework to develop a survey specifically related to the SEARCH context to assess the impact on organizations. The steps followed and progress include: (MINUTES - JANUARY 8 EVAL STEERING COMMITTEE)

The embedding of evaluation in their approach is seen in the following quote.

> SS gave a history of the evaluation process within the SEARCH Program, including the

> establishment of the Evaluation Steering Committee, the end result of which was the development of the Evaluation Blueprint.
>
> SS outlined the goals of this meeting:
>
> - To review key findings from completed and in-progress evaluation projects
> - To provide feedback on the development of the "Organizational Research Capacity Model", and
> - To identify implications and distil key recommendations for Steering Committee (and others) (MINUTES - JANUARY 8 EVAL STEERING COMMITTEE)

There is evidence that external expertise was also sought to move forward with the programme evaluation plans.

> A workshop involving seven experts (researchers in relevant areas and two practitioners) plus AHFMR staff was held for the purposes of developing a conceptual framework through which to begin to understand the capacity of an organization to create and use research knowledge. There were two variations of models developed. These models were shared with the group and feedback sought. (MINUTES - JANUARY 8 EVAL STEERING COMMITTEE)

There is also evidence that the Steering Committee set the priorities for this process.

> Discussion centered on what Steering Committee members would include as the priority goals of evaluation of SEARCH. Thoughts and opinions expressed included: (May 23 2001 minutes)

And that they examined the results.

> There's been follow-up in on the 14 recommendations outlined in the SEARCH Program Evaluation Blueprint, commissioned by the program to identify the primary questions that program stakeholders (particularly participants and their organizations) want answered? AA reviewed the Program's activities in response to each recommendation, as well as the plans for seven specific research/evaluation projects to answer priority questions. (as in attached action plan) (STEERING COMMITTEE March 7minutesfinal)

I think it is important to point out that the committees overseeing the SEARCH Program were active and continually questioned the direction of the programme and the impact that it was having.

> It was felt that both the individual and organization level evaluation pieces together will provide a more complete picture of SEARCH's impact. They also may lead to greater understanding about how one influence's the other.
>
> The Committee provided feedback and suggestions for additional analysis as well as overall implications for program design and delivery. (MINUTES - JANUARY 8 EVAL STEERING COMMITTEE)
>
> Evaluation in the future will need to answer different questions - we need clarity about how we know we are down the road. It should also include how and why has it worked to move us down the road. The result will guide us as well as others to reproduce and model what we've done. It looks at the process used to achieve the outcome. (May 23 2001minutes)

So the approach to evaluation was detailed, designed to be integrated and to span the long term of the programme. It demonstrated that there was a culture of evaluation embedded in all aspects of the planning and delivery of the SEARCH Program.

5.8.3 Programme innovation

The last data category to be presented relates to evidence that supports the premise that changes that were made in the delivery of the SEARCH Program were not minor 'tweaks' but substantive alterations as you would expect to see in an environment that was using a developmental evaluation approach. Although there are multiple examples, three areas have been chosen that clearly provide evidence to support this hypothesis; programme delivery (method and locations), role of faculty (including curriculum development) and use of technology. It is also worth noting that the data come from the qualitative data described above but has also been informed by the evaluations included in the quantitative data presented earlier.

Prior to presenting these data, it is worth noting one of the recommendations from the March, 2001 Steering Committee meeting (timing would be the end of SEARCH III cohort and recruitment of SEARCH IV).

> There is a need to keep overall SEARCH Program goals consistent throughout one iteration of SEARCH while recognizing the value of reviewing the goals regularly.

It was obviously recognised that there had been extensive changes made in the programme and more were planned but there was also a recognition that some stability was required.

Programme delivery

A significant change in the method of programme delivery took place over the history of the programme, with the greatest changes occurring during the first three cohorts. These changes were driven by two primary forces: the participant evaluations and the overall goal of the programme to develop working networks of health care professionals across the province.

The research that forms the basis of this report does not include the evaluation of module sessions. Students evaluated every session that was presented during the residential modules. This was done through an on-line evaluation form that was submitted at the end of the session and the results were emailed to the faculty within an hour of the completion of the session. These evaluations were examined at the end of each day and where necessary changes were made to the content on the following day. Therefore feedback was integrated into the following sessions and was also considered in the development of future modules and the programme overall.

Taken from a broader programme perspective the primary programme delivery changes were related to the module structure and timing and the location of the courses. Table 22 outlines the significant changes that occurred over the first four cohorts of the programme.

Table 22 Evolution in programme delivery and faculty over first four cohorts

Cohort	Delivery	Location	Faculty	Others
SEARCH I	Two seven week sessions in the first year although the programme ran over two years	Single (Banff)	Primarily visiting faculty	None noted
SEARCH II	7 x 1 week sessions spread over first 18 months of the two year programme	Single (Banff)	Core and visiting faculty	None noted
SEARCH III	7 x 1 week sessions spread over two years	Multiple around the province	Core faculty	Managers session CEO of local health authority Representatives of ethics committees
SEARCH IV	7 x 1 week sessions spread over two years	Multiple around the province	Core faculty	As above

As can be seen there were changes in the delivery mechanism. The first cohort spent significant blocks of time away at the course and then went back to work in their health authorities. This was particularly difficult for many participants as identified in the course evaluations, and a decision was made to spread the programme over a larger number of shorter modules. Having the modules in the resort of Banff was also viewed by colleagues (of participants and faculty) as a bit of a holiday in a resort location. I can attest from personal experience that these sessions were anything but a vacation. Examination of a standard schedule for the

course (Appendix 1) clearly demonstrates that the days were fully booked with taught sessions and evenings were used to develop individual and group projects.

The decision to move the module locations around the provincial regions was instigated by a northern health region. This increased the time and travel costs of the programme (AHFMR covered all travel and accommodation costs for faculty and participants). However, the strategy was consistent with the goal of developing provincial networks. Not only did participants get first-hand knowledge of the different health regions (e.g. there were presentations related to local initiatives), but members of those health regions also had an opportunity to become more familiar with SEARCH Program through contact with participants and faculty.

These evolutions in the programme demonstrate the use of ongoing evaluation to significantly alter the method of delivery of the programme in an attempt to meet programme aims in an evolving context.

Role of the faculty

As in noted in Table 22 the faculty delivering the programme changed significantly up to and including the SEARCH III cohort. The first cohort was taught primarily by external visiting faculty. By the time of the second cohort there were local faculty (some of whom became core faculty), but the balance was still in favour of external visiting faculty. Examination of the minutes identifies the Steering Committee's awareness of the dissatisfaction of the students due to the lack of cohesion and continuity in the delivery. A decision was taken to establish core faculty responsible for programme delivery, and external contracts were established with a number of universities and a private organisation.

Notes from the June, 2003 Steering Committee clearly outline the changes that took place over the first three SEARCH cohorts.

> Faculty Development
>
> AA provided background on the role and make-up of the Faculty during SEARCH since 1996:
>
> SEARCH I: The Foundation funded two faculty at 50% time - one in Edmonton and one in Calgary, and the others were visiting lecturers.

> SEARCH II: Continued with two consistent Faculty members, but increased involvement of local teams to develop and deliver specific modules. Faculty reported a sense of isolation, and lack of continuity or full engagement in program development.
>
> SEARCH III: The faculty support and engagement was re-designed with the current approach to supporting 10 people consistently across the full 3 years of a program plan and delivery, with different levels of commitment for Lead and Core team members. Also, established the 'theme teams' around the curriculum framework. Faculty report an increased sense of engagement and personal and professional satisfaction. We now have a very solid, supportive, and highly functional multi-disciplinary group, who are able to continuously develop and deliver the curriculum. (MINUTES OF JUNE 19 2003 STEERING COMMITTEE MEETING)

However, this did not herald the end of changes with regards to the programme. The core faculty, prompted by the continuous evaluation feedback spent the next five years redesigning and improving the core curriculum of the programme. So although the core themes of the curriculum remained unchanged there were significant changes to the ways which core elements were delivered, with a focus on integrating the teaching from each element and as noted earlier success in this was reported in this integration.

Use of technology

The SEARCH Program both benefitted and suffered from the use of technology. As noted in the introduction and demonstrated by the two papers published regarding the use of IT, the SEARCH Program was leading edge in what it provided for students (Lau and Hayward, 2000, Lau et al., 2001). The use of technology was co-ordinated through the Centre for Health Evidence (CHE) at the University of Alberta. This was part of a nationally funded programme and an overview was provided to the Steering Committee in January, 2001.

> DD spoke on the history of CHE - when it began three years ago, it was through the Office of the Health Information Highway, Health Infostructure Program of Health Canada with partnership funding including support from AHFMR. CHE Partners include: University of Alberta, Capital Health, Infoward, AHFMR, and Health Canada.
>
> The HIS Program was intended to show what an evidence based health information system could look like and brought government, public sector, private sector and health regions together. (Steering Committee Jan 01 Mins)

The first published paper by Lau and Hayward (2000) related to the use of technology provides a classic example of 'ready, fire, aim'. The paper uses four categories to describe the evolution of the technology through the two year SEARCH I cohort. The categories are; defining expectations, initial development, coping with technology and improvements over time and the paper presents the

activities that took place over one to three month periods of time . In the initial development category the paper reports;

> The program integrator was installed during the second training session in July. Shortly after its introduction, many software bugs were detected in it, which required immediate fixes by the developers. The complex configuration of different software on the note books and the support staff's lack of prior exposure to the integrator made it difficult to diagnose and correct many of the technical problems that occurred.(pg 366)

The paper also demonstrates the extensive data that was collected as part of the evaluation of the use of technology (see Table 23).

All the early evaluations identified issues with the use of the laptop computers both during the modules and later when participants returned to work. Many of the issues were due to the technology but were also a result of the participants' lack of familiarity with the technology. It was noted by early SEARCHers that they had better computers and better access to on-line resources than the majority of the staff that they worked with. It is not within the scope of this thesis to discuss the advanced nature of the IT services made available to the participants.

However, participants were working in a period of rapidly changing technology. There is evidence from the qualitative data that the platform used for the programme was constantly being updated and new facilities added to allow students greater access to external resources. This included the early adoption of what has come to be known as WiFi, which happened in the SEACH III cohort (2000) long before it was being commonly used in other settings.

Table 23 Types, volume and sources of data collected over two years

Type	Volume	Source
Program documents. These included pre-training surveys, computer instructional objectives, course outlines, technology feasibility study, project selection criteria, project milestone map, computer support policies, development of second training program.	10 sets of documents	Staff, organizers, coordinators, and participants. Given to researchers.
Participant interviews. Three sets of telephone interviews conducted in Dec 96, Jun 97 and May 98.	63 interviews	Participants. Collected by researchers.
Staff Interviews. Face-to-face interviews with project sponsor, coordinators, and support staff conducted in Dec 96, Apr 97, and Apr 98.	12 interviews	Staff. Collected by researchers.
Meetings. Notes from meetings with coordinators, curriculum subcommittee, technology and content support staff, and facilitation sessions.	34 meetings	Minutes recorded by staff; notes by researchers.
Online surveys. Automated online surveys from program integrator consisted of one set of registration surveys and three sets of interval surveys collected in Oct 96, Apr 97, and Apr 98.	46 surveys	Participants. Summarized by researchers.
Discussion groups. Computer discussion conferences were for participants and were moderated by participants.	16 conferences 14 surveys	Participants. Summarized by staff.
Program Web site. The Web site was maintained by program staff with 15 hypertext-linked sections and monthly Web site hit rate statistics	15 sections 19 months-hits	Participants, Web stats by staff. Given to researchers
Help desk logs. Logs recorded the history of technical assistance provided to participants and staff from Jul 96 to Jan 97.	267 log entries	Technical staff, participants. Given to researchers.
Computer usage. Three sets of application usage and online survey data from the program integrator of each participant's notebook were collected in Oct 96, Apr 97, and Apr 98.	30 sets of usage data	Participants. Collected by researchers.
Training courses. Workshops on resource inventory, needs assessment, grant proposal writing. Microsoft Access, and distance education used face-to-face meetings, an interactive Web site, and video conferencing	31 feedback 1 group input	Participants. Collected by staff; forwarded to researchers.
Program evaluation. Evaluation reports produced by independent consultants provided by participants, regional executives, and managers at seven weeks, six months, one year, and 18 months were evaluated by independent consultants.	4 reports	Collected by independent consultants.

Adapted from Lau and Hayward (2000)

An overview of the evolution was provided to the January 2001 Steering Committee.

> DD reported on the history of SEARCH in terms of informatics supports and technology. In SEARCH I, everyone got laptops; in SEARCH II, everyone had internet, with an on-line curriculum. AHFMR provided hardware and software. In SEARCH III, the Centre for Health Evidence, working with the Institute of Professional Development, has taken on the role of
>
> - Creating an on-line virtual community,

- Integrating knowledge resources required to conduct program

> DD commented that IT has developed to a point that makes managing a distributed community much easier: all participants need is access to the internet. There is great diversity among the RHA's in Alberta, and CHE has experience with brokering these discussions. There isn't going to be one answer. (Steering Committee Jan 01 Mins)

There is evidence of the conflict that arose when the health authorities began to establish their own IT systems (SEARCH IV), and instead of the SEARCH Program providing computers for participants the health authorities were given grants to purchase computers compatible with their internal systems.

> There was some discussion on the model that is being adopted this time for provision of hardware. In order to address issues that have occurred in the past, such as insurance, maintenance, and support in the RHA's, a "granting approach" is being taken. RHAs will be granted the money to purchase a computer according to SEARCH program specifications. The Committee members felt that this would be very advantageous to some of the regions. It was stressed, though, that involvement from IT in the regions was important, especially in matters of maintenance. (Steering Committee Jan 01 Mins)

What is clear is that the changes that were taking place in the use of technology were leading edge and continuously evolving to keep up with the new and available on-line resources (things we take for granted today) and that these changes had an impact on the programme and the role of the students in their work environments.

In summary the 'ready, fire, aim' approach that had been used to establish the SEARCH Program is clearly demonstrated through this analysis and also demonstrates that it continued throughout at least the first four cohorts of students. There is also evidence that the results of the evaluation led to extensive changes in terms of programme delivery, the role of faculty and the use of technology meant that evaluations were used to make dramatic changes to the programme over that period of time and that such changes would fit within a context of the use of developmental evaluation and not simple formative or summative evaluation.

The following chapter discusses the findings of the meta-evaluation and the qualitative data analysis.

6 DISCUSSION

The results section provides a description and analysis of the extensive evaluations that were carried out over the life of the SEARCH Program. It has demonstrated that these evaluations were not haphazard but were a part of an overall evaluation plan (Evaluation Blueprint) designed to provide feedback to programme funders, faculty, participants and health regions. The following discussion examines the extent to which the original research aims of this thesis have been achieved and the questions posed have been answered. The aim and research questions are presented here for reference.

Research aim

To critically examine and assess the applicability, use and practices associated with evaluation within the context of programme documentation and programme evaluations related to a continuing professional development programme for health care professionals.

Research questions

The specific research questions that guided the research were;

1. What was the quality of the SEARCH Program evaluations when assessed using international quantitative standards for programme evaluation?
2. What role did programme evaluations play in the development and evolution of the SEARCH Program?
3. What implications do these two perspectives have for the evaluation of future continuing professional development programmes?

This section will first discuss whether the selection of the case study proved to be appropriate to address the research aim and subsequent research questions. It then goes on to discuss the outcome, strengths and limitations of the use of the metaevaluation and specifically the usefulness of the Stufflebeam metaevaluation tool to retrospectively assess the evaluations that were carried out as part of the SEARCH Program. The following section examines the concepts of programme development in relation to the evaluation and their fit with the concepts of developmental evaluation. The chapter ends with implications for future CPD programme development and evaluation practices.

6.1 Use of case study method

One of the most critical decisions to be made after establishing your research questions is the choice of methods to address them. In this instance the research aim focused on a specific CPD programme and therefore the selection of a case study methodology seemed simple and straightforward.

However, Tight (2010) provides a differing perspective. He suggests that the use of the term 'case study' is too broad, has been poorly defined and therefore in the best case scenario, has a limited meaning, or in the worst case is misleading. In his recent paper he presents rationale to demonstrate that the use of 'a case study' be discontinued as a research method and replaced by what is actually carried out – for instance *' a detailed examination of or a detailed analysis of X'* (pg 338). Although I agree that there is confusion regarding what a case is, simply stating what was done has the potential to exclude important information. The most likely omission is in the exploration and definition of just what 'X' is. In addition within a case study, this one for example, there may be more than one analysis that is carried out. Let us deal first with the difficulties of defining just what is the case being studied.

The concept map as presented by Ragin (1992b) was introduced in the methods section and is presented here again for reference (Table 24).

Table 24 Conceptual map for cases

Understanding of cases	**Case conceptions**	
	Specific	**General**
As empirical units	Cases are found	Cases are objects
As theoretical constructs	Cases are made	Cases are conventions

Adapted from (Ragin, 1992b)

There is no question in my mind that at the beginning of the project the case I was dealing with was contained in a specific empirical unit, the SEARCH Program. I was not theoretically creating new concepts related to CPD for health care professionals, I was dealing with a real existing programme. However, the mapping suggests that the case may be considered specific or general. In relation

to this research the programme itself was the identified case but it could be argued that the evaluations and records were the objects as cases that would be considered in the second cell. Ragin (1992) goes on to point out that the lines between the categories are not solid and may become blurred. That is indeed what happened my research. My initial thoughts firmly placed this case study in the first cell with the SEARCH Program as the defined case. However, as discussed below after more in depth consideration this was no longer as clear as it seemed at first inspection.

Hammersley (2010) agrees and points out that the definition of the case might not be clear at the beginning of the research project and might change over time. This is consistent with the ideas presented in the conclusion to Ragin's book (1992) where he describes what he calls 'casing' which deals with this issue.

Ragin (1992) describes 'casing' as a methodological step that can take place at any phase of the research and refers to the thought processes and consideration of aspects of the case being studied that evolve as the case is examined and defined or re-defined over time. I found Ragin's discussion of casing complex, with multiple theoretical twists and turns. However, what the concepts of casing provided for me was encouragement to examine the evolution of thinking that brought me to my case and therefore to more clearly define what it was.

My casing process began with an overarching interest in the concepts of life-long learning and CPD for health care professionals. My thoughts were based on acceptance of the theories of principles of adult learning and also theories of change. Within theories of change the work of Lewin (1964) was particularly influential as its processes of field analysis can be applied to what adult learners need to consider if they are to use the newly gained knowledge from their CPD experiences in their work environment – in this case the implementation of EBP.

The next stage of casing process took me to the identification of various models of delivery of CPD. Having experienced a number of different models from the perspective of both being a student and a teacher I had developed a view of some of the strengths and weaknesses. This brought me to a specific model of delivery -

that used in the SEARCH Program - which was made up of residential modules with work-based application of learning. I was not aware of it at the time but in retrospect I would base the model within the concepts of communities of practice (Lave and Wenger, 1991, Wenger, 2006), although the complexity of the delivery system means that other theories could also be considered.

Now if my thinking had stopped there I would have been very clearly in the empirical/specific cell of the conceptual map. However, with my previous experience of various educational delivery models, and following discussions with SEARCH faculty, the next stage in my thinking took me to the importance of the evaluation of such programmes. At this point I was then in that grey area between having the SEARCH Program as the case or considering the evaluations and programme records as the case objects, all of which moved me into the empirical/general cell of the conceptual map with these documents as the objects forming the case.

However, given the complexities of Ragin's casing argument and Tight's admonition that the term 'case study' should not be used, I re-examined my position. To do this I relied on two sources. I went back to the work by Yin (Yin, 2009) where clear arguments are used to demonstrate the value of using a case study approach together with a variety of data collection methods to address specific research questions. In addition, I examined two companion papers by Tellis (1997a, 1997b) in which he clearly outlines the use of a case and shows that it can be examined from different perspectives. In light of his arguments, I found that I was comfortable with the concept that the SEARCH Program was the case and the programme evaluations and documents were simply the units of analysis for the case. I am therefore confident in the use of a case study and that the methods used in considering the case made it possible to address the research questions as posed.

The research was exploratory, it used a variety of data, viewed through differing lenses and did not attempt to compare the results of those examinations or to determine which was correct (Ryan, 2006). It was based on evaluation theory, which as explained earlier is in a state of being defined and is evolving. Being

pragmatic meant that I found my thinking most clearly fitted with what Christie and Alkin (2008) have labelled the 'use' branch of their evaluation theory tree and this led to the work of Stufflebean (1999, 2001b) using checklists and the newly evolving work of Patton (Patton, 2011) and developmental evaluation.

Given this background the following section reflects on the findings that were reported in the previous chapter.

6.2 Quality of SEARCH Program evaluations

Coming to this project with a background in systematic reviewing and meta-analysis, I was excited to discover that there were existing evaluation standards and concepts of metaevaluation (Stufflebeam, 1999, 2001b). As noted in the literature review, these standards were developed and updated through an extensive consensus process. They have been approved and adopted by the American National Standards Institute and this has provided the tool with a measure of content validity (Sanders, 1999).

Stufflebeam (Stufflebeam, 2001a) has presented a case for the use of checklists in evaluation and has headed a project to develop and promote their use. However, it is interesting to note that the standards used in his checklist have not been formally validated in practice. In a previous report Gould et al (Gould et al., 1995) report on the validation process that was used during the standard changes in 1994. The report indicates that the focus of the validation panel was the developmental process, the assumptions on which the standards were based, and the applicability of the standards. However, the majority of the report deals with the process of standards development, with only one small section (less than a page in a 25 page document) addressing the application of the standards. The report focuses on use in differing populations, not on validation of the standards, as would be expected given the mandate of the validation committee.

The applicability of the standards across various cultures has been addressed in the related literature. Russon (2000) brought together examples from a number of international programmes that examined the issue of the transferability to other cultures of American values included in the standards, and concluded that with

small changes they were still very valuable. The standards have also been used as a basis for the development of evaluation standards in a number of large international organisations such as UNICEF (UNICEF, 2004) and Danida (Danida's Evaluation Department, 2004).

One other report explored the use of the standards. It was a PhD project (overseen by Daniel Stufflebeam) that tested the correlation between ratings from different evaluators when the standards were applied to a pre-selected set of evaluations (Wingate, 2008). The project compared the assessments made by students, evaluation practitioners and evaluation scholars and found the corelation to be poor (Wingate, 2009).

Interestingly Cooksy and Caracelli (2005) report on a metaevaluation that used methods similar to those in this research project. Although they identify a number of evaluation standards/tools for assessing the quality of an evaluation (including the metaevaluation tool used in this project) they do not use any of them in their case study and instead choose to judge the quality of the evaluations in their case study using just two criteria: transparency of methods (including the clarity of the evaluation question), and the validity of the methodologies used.

Be that as it may, this project used the internationally accepted evaluation standards and to quantify the finding used the tool designed by Stufflebeam (1999). This analysis rated the evaluations as very poor with all of them failing to meet the basic pass requirement set by the tool's author. This was a disappointment but not a surprise, and it is worth examining the results of the quantitative analysis from three perspectives; the metaevaluation tool, the evaluation reports and the usefulness of quantitative analysis.

6.2.1 Metaevaluation tool

The first point to make about the tool is that the items and categories within it are not independent. For instance both submission and clarity of the report appear in more than one category and therefore are counted more than once. A closer examination of the items and categories identified a number of areas where such overlaps occurred (Table 25).

Table 25 Duplication of assessment items

Assessment item	**Assessment point** **Item number (factor number)**	**# of times counted**
Provision of report	U5 (1,2), U6 (2), P5 (all*), P6 (all), A1 (10), A3 (9)	25
Keeping stakeholders informed	U2 (5, 10), P1 (5), P3(3), P4(2), P6(3)	6
Provision of interim reports	U6 (1), U7 (6), F2 (6), P1 (8)	4
Meeting stakeholders' needs	U1 (8), U3 (1), U4 (5), A3 (1)	4
Training staff	F1 (5), A5 (5), A6 (6), A7 (2)	4
Hiring competent staff	U2 (1), F1 (4), P7 (1)	3
Minimising disruption	F1 (2), F3 (9), P4 (4)	3
Using independent evaluators	P7 (5), A1 (5), A3 (10)	3

* there are 10 items in each category

Given that all the evaluation reports relating to the SEARCH Program were clearly written, they then scored higher than they might otherwise have done if this had not been counted 25 times. Even so the overall scores for the evaluations were disappointingly low.

The use of the Stufflebeam (1999) quantitative formulae promised, I believe, a false sense of precision in the results. There is an assumption that you can add up the scores and at the end make a decision regarding the quality of the evaluation. Documentation regarding the tool does not provide a rationale for the components of the formulae, nor for the decisions regarding the selection of items that are so critically important that a score of poor means that the evaluation has failed (P1-Service Orientation, A5-Valid Information, A10-Justified Conclusions and A11 – Impartial Reporting). Using these criteria, all of the SEARCH Program evaluations failed to meet the minimum standard.

In an earlier publication Finn et al (1997) report the outcome of what they call a concurrent metaevaluation. They did not report the strict assignment of the scores used in the Stufflebeam checklist, which is interesting because Stufflebeam is a co-author of the report. Instead the standards are judged as falling into one of four categories: insufficient information, not met, partially met and met. Given that the more prescriptive assessment system was published in 1999, it is possible that the new system was, in part, a result of the previous work by Finn et al (1997). In

addition pass/fail criteria are not included in the 1997 report. The assessment of SEARCH Program evaluations in this thesis might have been somewhat more favourable if the broader categories from 1997 had been used but this possibility has not been explored. However, given the number of scores of 0 in the current analysis it is likely that the category of 'insufficient information' would have predominated.

It is possible that broadening the range of data collected and including interviews with SEARCH administrators, faculty and evaluators would have improved the scores. However, this was not possible in this project, because the evaluation tool was used retrospectively. There is a general problem with retrospective studies of this type, which raises questions, in the area of metaevaluation, about the their ability to adequately demonstrate the quality of the evaluations that have been undertaken.

However, there is even more doubt about the metaevaluation tool, given the interdependence of the assessment items, and the lack of clarity in the weightings used in quantitative scoring of the results. However, for the moment let us suspend judgement on that issue, and look at the quality of the research reports. It could be argued that using the tool in this retrospective manner was not the most appropriate approach and that the poor scores are a reflection on the content and quality of the reports, not the tool used to evaluate them.

6.2.2 Evaluation report quality

It is possible that the authors of the evaluation reports made assumptions regarding the knowledge and experience of their audience. This is almost certainly true of items in which the stakeholders were directly involved. The SEARCH Program had a long history of collaboration with stakeholders, and therefore the authors might not have felt a need to state explicitly in all of their reports that they identified and consulted them at the various stages of the evaluations. It is also well known that failure to report an activity does not necessarily mean that it was not done. Therefore there might have been any number of contacts between the evaluators and other interested parties that were not mentioned in their final reports.

Other omissions include two sets of missing data, one related to formal agreements/contracts, and the other to funding of the evaluations. AHFMR is a publicly funded institution and as such would be required to have appropriate contracts and audit processes in place. Other records in the archives indicated that contract letters did exist, and in the qualitative data analysis there were indications that formal roles for evaluators were defined.

It is also possible that I was overly demanding in terms of extracting the data in awarding a score of zero to items that did not specifically mention evaluation activities that were being scored. The benefit of the doubt was certainly not accorded to the evaluation report. Having said that, it is important to note that reports were assessed in total. For instance, when a number of individual reports from a specific evaluation were available they were grouped and evaluated as a single report. This allowed data to be counted if it appeared in the any of the interim or the final reports.

It was interesting to examine other reports of similar retrospective metaevaluation analyses and to find that their authors were able to address almost all the categories included when using a similar tool. Eichert (2008) reports a German retrospective metaevaluation of an organic farm programme. He found that it was not always possible to evaluate all the components on the metaevaluation checklist (DeGEval, 2008) owing to limitations within the project reports. In spite of this, the evaluation was rated very highly. Out of a total of 266 items, only 30 items were marked as impossible to evaluate and 45 were marked as unmet, with the remaining 191marked as met. Because the reports that were evaluated by Eichert (2008) are not accessible, it is impossible to ascertain why there is such a discrepancy between his reported use of the metaevaluation tool and my use of a similar tool in this project.

It is certain that the authors of the evaluation reports that I studies would be disappointed in the ratings that their reports received. As noted above, had it been possible to interview the evaluators to obtain further information, it is likely that scores would have been improved.

6.2.3 Usefulness of the quantitative analysis

However, the biggest disappointment resulting from the metaevaluation tool is that at the end of the process we only have a judgement of the quality of the evaluations. In this instance all the evaluations failed to pass minimum criteria. That does not necessarily mean that the evaluations were poor, or that they were not useful. As will be shown in the following sections, the evaluations were found to be very useful in the development and improvement of the SEARCH Program.

The metaevaluation tool does what it says on the box – it determines whether the evaluation process was of good quality. However, what is missing from the assessment is any comment on the overall outcome of the programme. This is equivalent to assessing the quality of a randomised controlled trial that evaluates a new therapeutic treatment as good, but neglecting to tell the reader that the new treatment was more harmful than the old one it was being compared to. Another example of the same approach would be a de-briefing on the handling of an emergency situation, outlining that health care professionals performed their jobs but failing to mention that the patient died. The focus is on the assessment of process when what is really important is the outcome.

There is a long history of improving health care using evidence about the quality of experimental studies. For example the CONSORT document (Moher et al., 2001) makes recommendations for assessing the quality of reporting of randomised controlled trials, and PRISMA for systematic reviews (Liberati et al., 2009, Moher et al., 2009). However the assessment does not stop at this point. In each of these areas endeavours have moved on, the aim being to combine data from a variety of research reports to assist policy makers in coming to conclusions regarding the efficacy and effectiveness of the interventions, and to make decisions of how those findings should/could be integrated into policy. This can be seen most clearly in the work of the Cochrane Collaboration (Cochrane Collaboration, 2011) in terms of assessing the effectiveness of health treatments, or in the UK the National Institute for Health and Clinical Excellence (2009), which assesses both clinical and cost effectiveness and in the development of MOOSE (Stroop et al., 2000) in the area of epidemiology

However, in the area of evaluation studies the combining of results from a variety of evaluations is relatively new, and raises important issues regarding which evaluations could or should be combined and the best methods for doing that (Farrington, 2003, Slavin, 2008). It needs to be acknowledged that there is a mechanism in place for reporting such research activities in the form of the C2 Campbell Collaboration (2010) which has been established to improve decision-making through systematic review in the area of education, crime, and justice and social welfare.

The metaevaluation process and the tool used in this project act as guidelines to judge the performance of the evaluator and but they lack focus on what is being evaluated. That is they examine how evaluators function – are they conducting high quality evaluations but do not touch on the important information about the programmes that they are evaluating. It is therefore useful as a tool for planning evaluations or cross checking the progress of the evaluation to ensure that the important aspects are managed as reported (Hanssen et al., 2008). It could also serve to provide the basis for the initial discussions between the evaluator and those commissioning them to identify the various perspectives of both.

In the context of this project using the metaevaluation tool to retrospectively judge the quality of an evaluation or a set of evaluations proved to be time consuming and did not provide particularly useful information in relation to either the evaluations or the programme being evaluated.

6.3 The developmental evaluation lens

In his introduction to Jamie Gamble's (2008) primer on developmental evaluation Michael Quinn Patton wrote;

> "....the answers will emerge from the process and won't be known until you engage in and reflect on the process. .....developmental evaluation will help you be clear about where you started, what forks in the road you took and why, what you learned along the way, and where you ended up, at least for a moment in time....." (pg 6)

In that primer Gamble provides background on the emergence of developmental evaluation through a set of workshops where Canadian volunteer organisations

met to address issues related to social innovations that they found difficult to evaluate. He describes developmental evaluation as embryonic with new ideas about it emerging all the time. The research reported in this thesis was an attempt to retrospectively examine the processes of the SEARCH Program to examine the role of evaluation in programme development and to do that through the lens of developmental evaluation.

6.3.1 Evaluation – a 'wicked' problem

A case could be made that the SEARCH Program was functioning within a complex environment, and that it was trying to address the needs of a number of different stakeholders. One could view this as being what has been described by Rittel and Webber(1973)as a 'wicked' problem.

Later Roberts (2000) outlined four characteristics of a 'wicked' problem;

- There is no definitive statement of the problem
- Different stakeholders therefore compete to frame the problem to their advantage
- The problem solving process is complex
- There are constraints as the problem definition and stakeholders change over time.

Conklin (2005) goes on to point out that solving wicked problems is not a linear process and that different people or groups will address the problem in different ways. He links wicked problem to social complexity. He also points out that attempts to solve 'wicked' problems frequently identify new 'wicked' problems as discussed by Roberts in his comments on the final characteristic in the above list.

It is worth examining whether this was the environment in which the SEARCH Program was founded. It came into being in the context of trying to implement research findings into clinical practice, which was a complex undertaking. Each group of stakeholders; health policy makers, health care practitioners, health care administrators and patients saw the problem in very different ways. Therefore addressing the problem was complex, and as actions were taken the situation changed and the issues needed to be re-defined. All of this firmly situates the implementation of EBP in the 'wicked' problem category. That is, complex, ill-

understood, evolutionary and changing over time, as the problem definition and stakeholders changed.

The SEARCH Program was established as a mechanism to address this 'wicked' problem. However, as previously discussed, the implementation of a possible solution simply presented us with another 'wicked' problem – how to determine whether the solution is working.

Roberts(2000) has suggested three different approaches to solving 'wicked' problems – authoritative, competitive and collaborative. The development of the SEARCH Program actually fell into both the first and the last of these categories. An authoritative decision was taken that there would be a CPD program and then collaborators were then identified to make it happen. In the words of Patton (2011) this was a 'ready, fire, aim' initiative. The approach to evaluation was the same. The SEARCH Program was established within an organisation that was focused on research and evaluation, and this evaluative culture was embedded in all aspects of programme activities.

6.3.2 Evaluative culture

The SEARCH Program functioned in an environment in which hard evidence was critical, and a culture of evaluation was embedded. It is worth taking some time to more clearly define what is meant by evaluative culture.

Mayne (2008) describes the characteristics of an evaluative culture, which are listed in Table 26. The data presented in the previous chapter demonstrated that the SEARCH Program, its faculty and administration exhibit these characteristics.

Table 26 Characteristics of an evaluative culture

Engages in self-reflection and self evaluation	• deliberately seeks evidence on what it is achieving, such as through monitoring and evaluation • uses results information to challenge and support what it is doing • values candour, challenge and genuine dialogue
Engages in evidence-based learning	• makes time to learn in a structured fashion, • learns from mistakes and weak performance • encourages knowledge sharing;
Encourages experimentation and change:	• supports deliberate risk taking • seeks out new ways of doing business

Adapted from Mayne (2008)

Mayne (2008) goes on to say that such a culture can be fostered through commitment from senior management, organisational support structures and an environment that has a focus on learning. Again, the SEARCH Program had the benefit of all of these. Examination of the programme records indicate that there was a shared understanding that evaluation was important and that all programme activities would be evaluated in some way and the results of such evaluations were consistently integrated into the development of the programme.

The data extracted as part of the qualitative analysis clearly demonstrated that he SEARCH Program was a complex evolving programme functioning in an even more complex and evolving health care system that was grappling with the difficulties of implementing EBP. In that situation, evaluation took place on micro and macro levels. In such situations standard formative and summative evaluation is not useful, because the problems are unbounded and it is likely that there are no right approaches – just some that are better than others.

It is important to note that having an evaluative culture as described by Mayne (2008) does not in and of itself mean that developmental evaluation is also taking place. Programmes can have a culture that includes on-going reflection, learning and experimentation but not be involved in developmental evaluation. It is therefore worthwhile to examine the activities with the SEARCH Program to ascertain their fit within what is coming to be known as developmental evaluation.

6.3.3 Fit with developmental evaluation

Taken individually, the evaluations carried out as part of the SEARCH Program could be viewed simply as examples of formative and at times summative assessments. However, examined through a broader lens it is clear that whether they realised it or not, the faculty and administration of the SEARCH Program were engaged in developmental evaluation, with a variety of individuals taking on a leading role at various times in the programme.

Examination of the programme documents revealed a culture that valued and supported reflection and evaluation at all levels. Results from evaluations were examined critically by the various steering committees and significant programme changes were made as a result of those critiques. Evidence in three areas of the data was: programme delivery, faculty, and use of technology. Evidence has been provided to substantiate that each of these underwent significant changes over the span of the programme and therefore contributed to the developmental changes in the programme.

It is interesting to note that when changes were made in the programme delivery and curriculum the SEARCH Program faculty were not restricted to working within the academic arena. That is, they did not seek formal accreditation for a Masters or PhD programme. If that road had been chosen, they would have been severely limited by institutional policies, and would not have been able to make the wholesale changes that they did in curriculum design and delivery. Instead, they remained independent and continued to focus on the work of participants in their work environment. In his discussion of the value of work-based learning, Garnett (2001) supports their decision as he points out that *'in the age of the "knowledge driven economy" and the "corporate university" the creation and evaluation of knowledge is now too important and all pervasive to be left to higher education'* (pg 78).

Although this was a decision that favoured the evolution of the programme, as noted in the discussion, the faculty were put in the difficult position of working for two masters, and their work with the SEARCH Program did not always fit in with the academic requirements of their universities. In addition, a lack of

ownership within the institutes of higher education made the programme vulnerable to the funding cuts that eventually caused the programme to be closed.

So in answer to the question whether the evaluative practices used in the SEARCH Program led to programme development and evolution, the answer is definitely yes. These practices have be situated within the context of developmental evaluation The evaluation processes used match those outlined above, in that they clearly demonstrated where the programme started, what roads were taken and why, what the participants and faculty learned along the way and where they were currently situated. Evaluation formed an intrinsic part of all activities, and the data demonstrate that substantive changes were made as a result of the findings of those evaluations

6.4 Implications

The third research question posed as part of this thesis asks how the findings might inform the evaluation of future CPD programmes.

The question of whether the conduct of developmental evaluation is achievable across a variety of situations is a very difficult question to answer. That is, how realistic it is to conduct developmental evaluation? The SEARCH Program was unique in at least three ways, which all made developmental evaluation possible.

The first was the level of support that it received from AHFMR. This support was not only financial but also provided the early and ongoing vision of how the programme fitted into the larger health care system of the province, and support for capacity building within that system, most especially in relation to the implementation of EPB initiatives.

The second was in the quality, commitment and experience of the faculty and evaluators that worked with the programme. As noted above, at various times various faculty members took on the role of developmental evaluator as they identified evaluation needs, reviewed evaluations, and considered how the programme could be changed and improved. They acknowledged the complexity of both the programme and the system in which it functioned, and were eager and

willing to work with both to provide what they believed to be an innovative and important programme.

The third was the link established with CHE and their collaboration which was a leader in the country in the introduction of the use of technology, and the collaboration that resulted from that link. This brought the programme into contact with individuals who were working on the leading edge of technology development, and meant that SEARCH participants had access to the most up-to-date technology.

Experienced evaluators will acknowledge that the convergence of such factors is wonderful when it happens, but it does not happen frequently. Evaluators are often faced with the problem of attempting to evaluate complex programmes with standard tools that are not up to the task. Gamble (2008) points out using such tools may not only be unhelpful but can actually be harmful if evaluation questions are too narrowly focused and therefore incorrect conclusions are drawn. Programme managers are often asked to provide causal links between their programmes and complex outcomes that are just not possible to demonstrate.

Therefore, no concrete recommendations are made with regard to the evaluation of future programmes beyond the obvious that it needs to be embedded in all programme activities and faculty and participants need to be reflective and open to change as dictated by the outcomes of the evaluation process.

This discussion of the findings of the research needs however to be considered with the limitations of the research in mind.

6.5 Limitations

6.5.1 Personal perspective bias

It would be irresponsible of me to conceal the particular perspective that I brought to this project. I have been both a student and an instructor in a number of different models of CPD for health care professionals. Each of the programmes had both positive and negative features. For example, there are benefits and

disadvantages in full-time or part-time study, integration of learning in the workplace environment, and the development (or not) of professional networks.

When I first encountered the SEARCH Program I was struck by the vision of its founder, the innovative educational model, and the level of evaluation being carried out even in the early stages of the programme. I became involved in the programme as a visiting faculty member during the SEARCH III cohort and I also collaborated with SEARCH faculty and participants at workshops held in Liverpool in 2005. The aim of the workshops was to gain interest in the NHS and the University of Liverpool with a view to developing a similar programme. The evolution of the research project that has formed the basis of this thesis and its focus on evaluation came through an iterative process of discussions with SEARCH Program faculty, discussions with my supervisors, and my continued desire to develop and deliver quality CPD programmes for health care professionals in the UK.

6.5.2 Data availability

Although the SEARCH Program began in 1996, electronic records were only available from 2000. However the existing data and records of the early evaluation activities provided clear reports of those evaluations and an outline of the evolution of the programme over time.

6.5.3 Selection bias

The initial list of completed evaluations was provided by the SEARCH Program administrator, and I selected the data sources to be extracted which have resulted in selection bias. However, this study was not meant to be comprehensive, but to provide information about the evaluative practices, and their possible fit within the contexts of metaevaluation and developmental evaluation. Both Dressman (2008) and Anyon (2009) point out that data do not speak for themselves, and that qualitative researchers find what they are looking for . That was certainly the case in this piece of research. However Dressman (2008) also points out that data are not discrete entities but part of the rich network to which they belong, and need to be interpreted in that context. I believe that my experience with the programme and faculty allowed me to do just that.

6.5.4 Data coding and analysis

I was the only quantitative data extractor, and although an early cross check was undertaken of intra-rater reliability, there was no comprehensive quantitative data checking mechanism in place. As noted earlier, although there most certainly are data extraction errors, given the overall poor quality rating of the evaluations such errors would have a limited impact.

In terms of the qualitative data, I was the single data coder and made all decisions regarding coding categories. This could be seen as a providing a significant bias in the management of the data. However, I was also familiar with the programme and the faculty and as such was able to link the context of the evaluations and the meeting minutes with what was happening with the programme at the time. For instance, having been present during SEARCH III and again toward the end of the programme I was able to link the data to events that were happening on the ground (e.g. changes in the health regions and the uncertainties this caused the programme).

Even with the perspective I bring to the examination of this data it is clear that anyone wanting to replicate the work could follow the data analysis plan as described. The one thing they would not bring to the process is my experience with the SEARCH Program and faculty. However, even without that I believe that they would come to similar conclusions from the quantitative data analysis and the qualitative data regarding the fact that the SEARCH Program was a complex and evolving programme working within a complex environment and that they had a continuous view to evaluating themselves and how they were functioning in and impacting on that environment. The matter of whether in fact this qualifies as developmental evaluation is more subjective issue.

7 CONCLUSIONS

In the context of continuing professional development for health care professionals, the SEARCH Program was an approach to learning that was innovative and arguably ahead of its time. It was an inter-disciplinary programme, the model of learning was collaborative, the method of delivery included classroom experience, mentorship and the integration of learning into the participants places of work. In addition to this it included the integration of the most up-to-date computer and internet technologies available at the time.

This case study was used to explore three research questions related to the evaluative practices of the SEARCH Program. The approach to answering these questions included the exploration of the programme evaluations and documents from two different perspectives. Although these perspectives are within the 'use' branch of evaluation theory they represent two very different (qualitative and quantitative) approaches to examine programme evaluation.

To answer the first research question regarding the quality of the evaluations two frameworks that have been used to guide programme evaluation and an internationally accepted set of evaluation criteria were used. RUFDATA and Impact are two frameworks that allowed for consistency in the extraction of information from the evaluations to allow for consideration of the content and context of the evaluations conducted during the various aspects of the evaluation process. RUFDATA demonstrated the integration of evaluation activities throughout the programme and also allowed for comparison across evaluations in relation to evaluation methods, uses, audiences etc. The use of the categories in the Impact framework allowed for the examination of the range of impacts measured during the extensive programme evaluation. In combination the frameworks allowed for a structured examination of the evaluations that had been conducted during the life of the SEARCH Programme.

The use of the nationally accepted standards for evaluation of programme evaluations, applied retrospectively, proved less than ideal or useful. This case study identified serious deficiencies in the metaevaluation tool designed to examine evaluations activities. Examination of the checklist identified that a

number of the items were interdependent, that is they measured the same factors on more than one occasion (e.g. the submission of the evaluation report was account for by 25 different criteria points). This, in principle, should have raised the scores for the evaluations considered, but in practice did not. The use of the checklist resulted in very poor scores for all of the evaluation reports.

The apparent precision offered by the metaevaluation tool and the formulae used to assess the results are also questionable. There is a lack of explanation provided regarding the weighting of the various items, the quantitative formulae used, and the criteria for classing an evaluation as a failure. It could be argued that contact with the author of the formulae might have provided this information. However, given that the evaluation criteria and the formula were readily available on the association website and their use promoted by the association, one would expect such information to also be readily available. In addition, the application of the standards was an intensive and time-consuming process

Although the metaevaluation tool has been developed using an extensive consensus process and adopted by international organisations, it has not been extensively evaluated. The only identified research report related to it demonstrated that there was poor reproducibility and correlation between different assessors when it was used. This assessment, although limited, included comparison of results of students, experienced evaluators and evaluation theorists. It is somewhat surprising that the checklist, which has been in use for more than 30 years has had such a limited amount of research done to provide evidence of its validity. It can only be assumed that this situation has arisen because of the intense consensus process that has been used to develop and amend the evaluation criteria.

Putting these reservations aside, reasons for the poor rating of the evaluations have been explored. It could be that the evaluators indeed did not conduct all the recommended evaluation activities recommended as part of the guidelines. Alternatively it could be that such activities were carried out but not reported in the evaluation reports. The fact that although the two primary evaluators were initially external to the SEARCH Program, their continued involvement with the

programme evaluation meant that as the years went on they were actually very familiar with the programme, the faculty and the participants and could have been considered insiders. As such there may have been information (e.g. the identification and inclusion of all stakeholders) that was so ingrained in the evaluation process that it was assumed and not reported in the evaluation reports. Also, criteria such as the details of the contracts between the SEARCH Program and the evaluators were not available. It could be argued that this is not information that would be included in an evaluation report but may be held in a different location and therefore retrospectively examining evaluations would not allow access to such information.

However, the greatest limitation of the tool is that it is focused on the evaluation process itself, and does not include any assessment of the merit or worth of the program being evaluated; nor does it provide a mechanism for the synthesis of different evaluations of the same or similar programmes. These are two very important limitations.

It appears that the development of the evaluation tool itself had its impetus in an effort to improve the quality of programme evaluations, through the identification of critical evaluation activities, as well as a tool to assess evaluations that had been carried out. As such the criteria were established through extensive consensus processes and it can be argued that using the criteria to plan and evaluate a given evaluation is valid and potentially useful. However, that still only provides information about the evaluation and tells the reader very little about the programme that has been evaluated. Such conclusions are important to the programme planners and are also important if attempts to bring together the findings from various programmes to look at their overall effects.

Mechanisms for such synthesis are now being developed within the C2 Campbell Collaboration (2010) which has been established to improve decision-making through systematic review in the areas of education, crime and justice, and social welfare. Until such syntheses are conducted, there will be little hope that the findings of this approach to evaluative research will be useful, as advocated by

Saunders, Trowler and Bamber (2011) or indeed that they will be able to inform the development of policy as called for by Pawson (2001).

In conclusion, the use of the RUFDATA and Impact frameworks allowed for consistent examination of and comparison across the evaluations. However, the evaluation criteria and tool used to judge the quality of the evaluations did not provide particularly useful information regarding the quality of the evaluations and provided no information about the quality of the programme being evaluated.

The second research question that guided this research related to the role of programme evaluations in the development and evolution of the SEARCH Program. The lens of developmental evaluation was used and it provided a more comprehensive overview of the evaluation process and the changes made to the programme being evaluated. Since developmental evaluation is a relatively new field, this case study provides evidence of the use and usefulness of this evaluation process in the specific context of an innovative and evolving continuing professional development programme.

The developmental evaluation lens was used to examine three key areas of the SEARCH Program: the environment in which the programme function, the evaluative culture of the programme and programme innovations. The first two of these areas demonstrated the complexity of the environments in which the programme functioned and the third highlighted the changes that took place within the programme as a result of the evaluations that were conducted. This examination identified that the evaluation processes used within the SEARCH Program were part of an overall evaluative culture and that responses to the evaluations were not in the form of minor adjustments to the programme but involved wholesale changes to the programme in relation to programme delivery, the role of the faculty and use of technology.

A case has been made that the provision of CPD programmes for health care professionals takes place in a complex environment and therefore programmes need to be innovative, responsive and flexible. In addition, given the current status of both the health services and higher education in the UK, such programmes need

to include collaborations across the sectors. Consequently, this area of development and evaluation needs to be considered as a 'wicked' problem.

Assuming that these programmes continue to evolve then the use of developmental evaluation might be helpful. However, it is important to point out that the criteria for the use of developmental evaluation need to be examined and compared to individual programmes to ensure that its use is appropriate to the situation. It is also acknowledged that the conduct of developmental evaluation is time consuming and requires commitment on the part of all programme collaborators. That is not to say that these attributes are not required in all good evaluation but are especially important when it is acknowledged that programmes may undergo dramatic changes during the evaluation process.

Therefore such development and evaluation might be difficult and will be challenging. However, SEARCH Program activities have demonstrated that this can be rewarding to both faculty and participants. It has also been shown through these evaluations, that it is unlikely that mechanisms can be established to measure, in a positivist manner any direct benefit for patients or the health care system. However, such evaluations can identify associations between the programme and the impact made within the work environment.

It is worth noting other conclusions and implications that conducting this research has highlighted and that link to the third research question. As noted at the beginning of this thesis this research was the result of a personal journey, a journey that continues and will include for me, the development of CPD programmes for health care professionals.

The complexity of implementing evidence into clinical practice in the UK has not become any easier or less complex with time – even with the guidance currently provided through the National Institute of Health and Clinical Excellence (National Institute for Health and Clinical Excellence, 2009). Decisions still need to be made regarding how best to implement such guidance and although the education of current basic health care providers has improved through the creation, choice and use of evidence, there is still a gap in the knowledge of these

practitioners. There is also a disconnect between the practice needs of these health care practitioners and the drivers of the academic research agenda. These gaps can be filled, but the research reported here identified that the mechanisms used to fill them will need to be flexible and responsive to the requirements of both the learners and the environments within which they work.

Extensive, critical review of the SEARCH Program model has only strengthened my earlier opinion of its merit and worth in terms of allowing for this flexibility and responsiveness. Both the SEARCH Program and INCLEN, on which it was originally based, were innovative in design and delivery and established outside of the standard academic environment – that is they did not offer academic credentials. They also had in common significant visionary leadership and financial backing. These factors allowed the programmes to develop and expand.

In addition, the SEARCH Program provided an environment that attracted faculty members who were both critical and creative thinkers and open to innovation and change in the design and delivery of the programme. This, in combination with the leadership and support from AHFMR helped create an evaluative culture which was necessary to allow for the developmental evaluation that took place within the programme.

Like Garnett (2001) I believe, that given the importance of the issue of CPD for health care professionals there is a need to provide learning opportunities that link directly to the workplace environment. Although there are exceptions, the lack of flexibility in standard academic programmes means that this is unlikely to happen within the current constraints of academically accredited programmes.

So those are the messages that take I personally from this research and will use in the development and evaluation of future CPD programmes for health care professionals.

8 REFERENCES

Alberta Heritage Foundation for Medical Research 2004. Fourth International Board of Review. Edmonton: AHFMR.

Alkin M & Christie C 2004. An evaluation theory tree. *In:* Alkin M. (ed.) *Evaluation roots: tracing theorists' views and influence.* Thousand Oaks: Sage. 12-65.

Anderson R. 2007. *Thematic content analysis: descriptive presentation of qualitative data* [Online]. Available: http://www.wellknowingconsulting.org/publications/pdfs/ThematicContentAnalysis.pdf [Accessed February 21 2011].

Anyon J (ed.) 2009. *Educational research: toward critical social explanation,* New York: Routlage.

Birdsell J & Mathias S 2001. 0109_SEARCH program evaluation, research and development blueprint. Edmonton: On Management Ltd 1-69.

Birdsell J, Zerbe W, O'Connell P, Thornley R & Hayward S 2005. Building capacity for evidence-based management in Regional Health Authorities. *In:* Casebeer A., Harrison & Marks (eds.) *Innovations in health care: A reality check.*

Brannen J. 2005. *Mixed methods research: a discussion paper* [Online]. ESRC National Centr for Research Methods. Available: http://www.ncrm.ac.uk/publications/documents/MethodsReviewPaperNCRM-005.pdf [Accessed April 2011].

Byrne D 2009. Case-based methods: Why we need then; what they are; how to do them. *In:* Byrne D. & Ragin C. (eds.) *The SAGE handbook of case-based methods.* London: Sage Publications.

Campbell Collaboration. 2010. *Campbell Collaboration Home Page* [Online]. Available: http://www.campbellcollaboration.org/ [Accessed January 2010].

Chalmers I & Altman D (eds.) 1995. *Systematic reviews,* London: BMJ Publishig Publishing Group.

Chelimsky E 1997. The coming transformations in evaluations. *In:* Chelimsky E. & Shadish W. (eds.) *Evaluation for the 21st century: a handbook.* London: Sage Publications. 1-12.

Chelimsky E & Shadish W (eds.) 1997. *Evaluation for the 21st century: a handbook,* London: Sage Publications.

Christie C & Alkin M 2008. Evaluation theory tree re-examined. *Studies in Educational Evaluation,* 34**,** 131-135.

Cochrane Collaboration. 2010. *Effective Practice and Organisation of Care Group* [Online]. Available: http://epoc.cochrane.org/ [Accessed April 2010].

Cochrane Collaboration. 2011. *Cochrane Collaboration Home Page* [Online]. Available: http://www.cochrane.org/ [Accessed April 2010].

Conklin J 2005. Wicked problems and social complexity. *In:* Conklin J. (ed.) *Dialogue mapping: Building shared understanding of wicked problems.* John Wiley &Sons.

Cooksy L & Caracelli V 2005. Quality, context, and use: issues in achieving the goals of metaevaluation. *American Journal of Evaluation,* 26**,** 31-42.

Cronbach LJ, Ambron S, Dornbusch S, Hess R, Hornick R, Philips D, *et al* 1980. *Toward reform of program evaluation,* San Francisco, Jossey-Bass Publishers.

Danida's Evaluation Department. 2004. *Meta-evaluation: private and business sector development interventions* [Online]. Available: http://www.um.dk/NR/rdonlyres/8784EC72-54D8-484B-8F8C-B1CC9791163A/0/Meta.pdf [Accessed October 2010].

DeGEval. 2008. *Evaluation standards (DeGEval-Standards), Germany* [Online]. Cologne, Germany: DeGEval – Gesellschaft für Evaluation (Evaluation Society). Available: http://www.degeval.de/calimero/tools/proxy.php?id=19084 [Accessed 2010].

Dickson R 2005. Systematic reviews. *In:* Hamer S. & Collinson G. (eds.) *Achieving evidence-basted practice*. London: Balliere Tindal. 43-62.

Dressman M 2008. *Using social theory in educational research: a practical guide*, Routlage.

Dwan K. 2010. *Within study selective reporting bias in meta-analysis* PhD, University of Liverpool.

Eichert C 2008. Meta-evaluation of action plans - the case of German Federal Organic Farming Scheme. *16th IFOAM Organic World Congress.* Modena, Italy: http://orgprints.org/view/projects/conference.html.

Farrington D 2003. Methodological quality standards for evaluation research. *The Annals of the American Academy of Political and Social Science,* 587**,** 49-68.

Finn C, Stevens F, Stufflebeam D & Walberg H 1997. The New York City Public Schools Integrated Learning Systems Project: a meta-evaluation. *International Journal of Educational Research,* 27**,** 159-174.

Flyvbjerg B 2006. Five misunderstandings about case-study research. *Qualitative Inquire,* 12**,** 219-245.

Gamble J 2008. A developmental evaluation primer. Ottawa: The J.W. McConnell Family Foundation http://www.mcconnellfoundation.ca/assets/Media%20Library/Publications/A%20Developmental%20Evaluation%20Primer%20-%20EN.pdf.

Garnett J 2001. Work based learning an the intellectual capitol of univerities and employers. *The Learning Organization,* 8**,** 78-81.

Gerring J 2007. *Case study research; principles and practices,* Cambridge, Cambridge Press.

Glass G 1976. Primary, seocndary, and meta-analysis research. *Educational Researcher,* 5**,** 3-8.

Glass G 1977. Integrating findings: the meta-analysis research. *Review of Educational Reserch,* 5**,** 351-379.

Glass G & Smith M 1981. *Meta-analysis in social science,* London, Sage Publications.

Glasser B & Strauss A 1997. *The discovery of grounded theory,* Chicago, Aldine.

Goldstein J 2008. Complexity science applied to innovation-theory meets praxis. *The Innovation Journal,* 13**,** 2-16.

Gould R, Basarab D, McGuire C, Robinson P, LeRoy F & Wigdor A. 1995. *The development, validation, and applicability of "The Program Evaluation Standards: how to assess evaluations of educational Programs."* [Online]. Joint Committee on Standards for Education Evaluation. Available: http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED403314&ERICExtSearch_SearchType_0=no&accno=ED403314 [Accessed April 2011].

Hamer S & Collinson G (eds.) 2005. *Achieving evidence-based practice,* London: Bailliere Tindal.

Hammersley M 2002. *Educational research: policymaking and practice,* London, Paul Chapman Publishing.

Hammersley M 2010. Unreflective practice? case study and the problem of theoretical inference. *Higher Education Close Up 5.* University of Lancaster.

Hanssen C, Lawrenz F & Dunet D 2008. Concurrent meta-evalution. *American Journal of Evaluation,* 29**,** 572-582.

Hayward S 2003. SEARCH evaluations overview. Edmonton: Alberta Heritage Foundation for Medical Research.

International Clinical Epidemiology Network. 2010. *INCLEN Inc* [Online]. INCLEN. Available: http://www.inclen.org/ [Accessed 2010].
Joint Committee on Standards for Educational Evaluation 1994. *The program evaluation standards*, Thousand Oaks CA, Sage.
Klein P. 2007. *Method versus methodology* [Online]. Available: http://organizationsandmarkets.com/2007/04/07/method-versus-methodology/ [Accessed April 2011].
Krippendorff K & Bock MA (eds.) 2009. *The content analysis reader*, London: Sage.
Langermann EC 1989. The plural worlds of educational research. *History of Education Quarterly,* 29**,** 185-214.
Langermann EC 1997. Contested terrain: a history of education research in the United States, 1890-1990. *Educational Researcher,* December**,** 5-15.
Lau F & Hayward R 2000. Building a virtual network in a community health research training program. *Journal of the American Medical Informatics Association,* 7**,** 361-377.
Lau F, Straub D & Hayward R 2001. Fostering an internet-based work group in a community health through action research. *Journal of Health Informatics Management,* 15**,** 207-221.
Lave J & Wenger E 1991. *Situated learning: legitimate peripheral participation,* Cambridge, Cambridge University Press.
Lewin K 1964. *In:* Cartwright D. (ed.) *Field theory in social science: selected theoretical papers.* New York: Harper Torchbooks.
Liberati A, Altman D, Tetzlaff J, Mulrow C, Gotzsche P, Ioannidis J, *et al* 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health cvare interventions: explanation and elaboration. *Annals of Internal Medicine,* 151**,** W65-94.
Madaus G & Kellaghan T 2000. Models, metaphors, and definitions in evaluation. *In:* Stufflebeam D., Madaus G. & Kellaghan T. (eds.) *Evaluation models; viewpointes on educational and human services evaluation.* 2nd ed. Boston: Kulwer Academic Publishers. 19-33.
Madaus G & Stufflebeam D 2000. Program evaluation: a historical overview. *In:* Stufflebeam D., Madaus G. & Kellaghan T. (eds.) *Evaluation models: viewpoints on educational and human services evaluation.* 2nd ed. Boston: Kluwer Academic Pulblishers. 3-18.
Magno C 2009. A metevaluaiton study on the assessment of teacher performance in an assessment center in the Philippines. *International Journal of Educational and Psychological Assessmenty,* 3**,** 75-93.
Mayne J 2008. Building an evaluative culture for effective evaluation and results management. *ILAC Working Paper 8.* Rome, Italy: Intitutional Learning and Change http://ageconsearch.umn.edu/bitstream/52535/2/ILAC_WorkingPaper_No8_EvaluativeCulture_Mayne.pdf.
Moher D, Liberati A, Tetzlaff J, Altman D & Group P 2009. Preferred reporting items for systematic reviews and meta-analysis: the PRISMA statment. *Annals of Internal Medicine,* 151**,** 264-269.
Moher D, Schulz K, Altman D & for the CONSORT Group 2001. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA,* 285**,** 1987-1991.
Muir Gray J 1997. *Evidence-based health care. How to make health policy and management decisions,* London, Churchill Livingstone.
Murphy C, Cross C & McGuire D 2006. The motivation of nurses to particpate in continuing professional education in Ireland. *Journal of European Industrial Training,* 30**,** 365-384.

National Institute for Health and Clinical Excellence. 2009. *NICE guidance* [Online]. Available: http://www.nice.org.uk/ [Accessed January 2010].
Nursing in Practice. 2010. *CPD Zone* [Online]. London. [Accessed 2010 Nursing in Practice].
Nutley S, Jung T & Walter I 2008. The many forms of research-informed practice: a framework for mapping diversity. *Cambridge Journal of Education,* 38**,** 53-71.
Patton M 2008. *Utilization-focused evaluation,* Los Angeles, Sage.
Patton M 2011. *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use* New York, Sage Publishers.
Pawson R 2001. Evidience based policy: I In search of a method - Working Paper 3. London: ESRC UK Centre for Evidence Based Policy and Practice http://www.kcl.ac.uk/content/1/c6/03/45/88/wp3.pdf.
Pawson R & Tilley N 1997. *Realistic evaluation,* Los Angeles, Sage.
Peters T & Waterman R 1982. *In search of excellence: lessons from America's best run companies,* New York, Warner Books Inc.
Platt J 1992. Case study in American methodological thought. *Current Sociology,* 40**,** 17-48.
Ragin C 1992a. "Casing" and the process of social inquiry. *In:* Ragin C. & Becker H. (eds.) *What is a case? Exploring foundations of social inquiry*. Cambridge: Cambridge University Press. 217-226.
Ragin C 1992b. Introduction: cases of "what is a case?". *In:* Ragin C. & Becker H. (eds.) *What is a case? Exploring foundations of social inquiry.* Cambridge: Cambridge University Press. 1-18.
Ragin C & Becker H (eds.) 1992. *What is a case? Exploring foundations of social inquiry,* Cambridge: Cambridge University Press.
Reese W 1999. What history teaches bout the impact of educational research in practice. *Review of Research in Education,* 24**,** 1-19.
Reynolds C. 2006. *A metaevaluation of NGO evaluations conducted under the AusAID NGO cooperation Program* [Online]. Available: http://www.ausaid.gov.au/publications/pdf/ngo_eval.pdf [Accessed October 2010].
Rittel H & Webber M 1973. Dilemmas in a general theory of planning. *Policy Scienes,* 4**,** 155-169.
Robert P & Engelhardt A. 2009. *OCHA Meta-evaluation - final report* [Online]. Available: www.lotus-group.org [Accessed June 2010].
Roberts N 2000. Wicked problems and network approaches to resolution. *International Journal of Public Management Review,* 1**,** 1-19.
Russon C. 2000. *The Progam Evaluation Standards in international settings* [Online]. West Michegan University: The Evaluation Centre. Available: http://www.ioce.net/download/reports/ProgEvalStandards-Intl.pdf [Accessed May 2011].
Ryan AB 2006. Post-positivist approaches to research. *In:* Antonesa M., Fallon H., Ryan A. B., Ryan A. & Walsh T. (eds.) *Researching and writing your theisis.* Maynooth.
Sackett DL, Richardonson S, Rosenberg W & Haynes RB 1997. *Evidence-based medicine. how to practice and teach EBM,* London, Churchill Livingston.
Sanders J 1999. General background on the Joint Committee on the Standards for Educational Evaluation. *Annual Meeting of the National Council on Measurement in Education.* Montreal, Canada: http://www.jcsee.org/wp-content/uploads/2009/09/JCGeneralBackground.pdf.
Saunders M 2000. Beginning an evaluation with RUFDATA: theorising a practical approach to evaluation planning. *Evaluation,* 6**,** 7-21.

Saunders M 2007. Widening participation capacity building in evaluation: Interim HEFCE report. Lancaster: University of Lancaster.
Saunders M, Trowler P & Bamber V (eds.) 2011. *Reconceptulising evaluation in higher education,* Maidenhead, England: McGraw-Hill.
Scriven M 1969. An introduction to meta-evaluation. *Educational Products Report,* 2**,** 36-38.
Scriven M 1997. Truth and objectivity in evaluation. *In:* Chelimsky E. & Shadish W. (eds.) *Evaluation for the 21st century: a handbook.* London: Sage Publications.
Scriven M. 2005. *Evaluation cafe: theory-free evaluation* [Online]. Kalamazoo: The Evaluation Centre - Western Michigan University. Available: http://vimeo.com/9409490 [Accessed December 2010].
SEARCH Canada. 2007. *SEARCH Canada History* [Online]. Available: http://www.searchca.net/users/folder.asp?FolderID=1382 [Accessed August 2011].
SEARCH Canada 2008. Accounting background operational and managment policies and procedures. Edmonton: SEARCH Canada.
Shadish W, Cook T & Levitan L 1991. *Foundations of program evaluation theories of practice,* Newbury Park, CA, Sage Publications.
Shadish W & Luellen J 2004. Donald Campbell: The accidental evaluator. *In:* Alkin M. & Christie C. (eds.) *Evaluation roots: tracing theorists' views and influence.* Thousand Oaks: Sage. 80-87.
Simons H 2009. *Case study research in practice,* Los Angeles, Sage Publications.
Slavin R 2008. What works? Issues in synthesizing educational program evaluations. *Educational Researcher,* 31**,** 5-14.
Smith M & Glass G 1977. Meta-analysis of psychotherapy outcome studies. *American Psychologist,* 32**,** 752-760.
Stacey R 2002. *Strategic management and organisational dynamics: the challenge of complexity,* Harlow, Prentice Hall.
Stake R 1997. Advocacy in evaluation: a necessary evil? *In:* Chelimsky E. & Shadish W. (eds.) *Evaluation for the 21st century: a handbook.* London: Sage Publications.
Stroop D, Berlin D, Morton S, Olkin I, Williamson G, Group ftM-aoOSiEM, *et al* 2000. Meta-analysis of observational studies in epidemiology; A proposal for reporting. *JAMA,* 285**,** 2008-2012.
Stufflebeam D 1974. Meta-evaluation. *Occasional Paper Series, Paper No. 3.* Kalamazoo, MI: Western Michigan University Evaluation Center.
Stufflebeam D. 1999. *Program evaluations metaevaluation checklist* [Online]. West Michigan University. Available: www.wmich.edu/evalctr/checklist [Accessed January 2010].
Stufflebeam D 2000a. Foundational models for 21st century program evaluation. *In:* Stufflebeam D., Madaus G. & Kellaghan T. (eds.) *Evaluation models: viewpoints on educational and human services evaluation.* 2nd ed. Boston: Kluwer Academic Publishers. 33-84.
Stufflebeam D 2000b. The methodology of metaevaluation. *In:* Stufflebeam D., Madaus G. & Kellaghan T. (eds.) *Evaluation models. viewpoints on educational and human services evaluation.* Boston: Kluwer Academic Publishers. 457-472.
Stufflebeam D 2001a. Evaluating checklists: practical tools for guiding and judging evaluations. *American Journal of Evaluation,* 22**,** 71-79.
Stufflebeam D 2001b. The metaevaluation imperative. *American Journal of Evaluation,* 22**,** 183-209.
Stufflebeam D, Madaus G & Kellaghan T (eds.) 2000. *Evaluation models. viewpoints on educational and human services evaluation,* Boston: Kluwer Academic Publishers.

Tellis W. 1997a. *Application of case study methodoloty* [Online]. Available: http://www.nova.edu/ssss/QR/QR3-3/tellis2.html#tellis [Accessed April 2011].
Tellis W. 1997b. *Introduction to case study* [Online]. Available: http://www.nova.edu/ssss/QR/QR3-3/tellis2.html#tellis [Accessed April 2011].
Tight M 2010. The curious case of case study: a viewpoint. *International Journal of Social Reasearch Methodology*, 13**,** 329-339.
Trowler P 2010. Wicked issues in situating theory in close up research. *Higher Education Close Up 5.* Unniversity of Lancaster.
UNICEF. 2004. *UNICEF evaluation report standards* [Online]. New York: UNICEF. Available: http://www.unicef.org/evaldatabase/files/UNICEF_Eval_Report_Standards.pdf [Accessed May 2011].
Wenger E. 2006. *Communities of practice: a brief introduction* [Online]. Available: http://www.ewenger.com/theory/index.htm [Accessed March 2010].
Wingate L. 2008. *The new metaevaluation standards: metaevaluation joins utility, feasibility, propriety and accuracy* [Online]. Kalamzoo: The Evaluation Centre - Western Michegan Universy. Available: http://vimeo.com/7560371 [Accessed April 2011].
Wingate L. 2009. *The program evaluation standards applied for metaevaluation purposes: Investigating interrater reliability and implications for use.* PhD, Western Michigan University.
Yarbrough D, Shulla L, Hopson R & Caruthers F 2011. *The program evaluation standards: a guide for evaluators and evaluation users,* Los Angeles, Sage Publications.
Yin R 1984. *Case study research: design and methods,* London, Sage Publishers.
Yin R 2000. Case study evaluations: a decade of progress? *In:* Stufflebeam D., Madaus G. & Kellaghan T. (eds.) *Evaluation models. viewpoints on educational and human services evaluation.* 2nd ed. Boston: Kluwer Academic Publishers.
Yin R 2009. *Case study research: design and methods,* London, Sage Publishers.
Zhang Y & Wildemuth B. 2009. *Qualitative analysis of content* [Online]. Available: http://www.ils.unc.edu/~yanz/Content_analysis.pdf [Accessed 21 February 2011].

9 APPENDICES

Appendix 1 SEARCH Curriculum Overview

Curriculum Themes

The SEARCH Classic curriculum is divided into three distinct, but interconnected theme areas: Creating Evidence, Choosing Evidence and Using Evidence.

Within each theme are a number of ‘threads’, that when woven together, result in a tightly integrated curriculum that teaches important skills and techniques related to applied health research and evidence-based decision-making. These themes are, where possible, taught within the context of the health environment in which the participants work. The theme areas and individual threads within each theme are presented below.

Each theme is taught by faculty members with expertise within that area. All three themes overlap and integration of the whole curriculum is ensured. Joint teaching occurs where ever possible to address common topics and issues.

Creating Evidence: research paradigms, policy and process; research designs, methods and techniques; evaluation and assessment methods; sources, analysis and management of health data; research ethics; research proposals, writing and presentation.

Choosing Evidence: information skills; health information systems; health knowledge sources; information searching and retrieval; critical appraisal of research studies; evidence-based guidelines; research synthesis.

Using Evidence: team work and collaboration; organizational change and change management; managing the interface of research and practice; health policy issues and evaluation; decision making; dissemination and communication.

The following is an illustration of the SEARCH Classic Curriculum Framework:

SEARCH III Module 7: Curriculum

Time (unless otherwise stated)	**Sunday November 3**	**Monday November 4**	**Tuesday November 5**	**Wednesday November 6**	**Thursday November 7**	**Friday November 8**
Morning 8:30am – 12:00pm		**From the Front Lines: SEARCH I and II on Dissemination in Real Life** *Sharon Matthias* **Dissemination: Developing a Plan for Sharing Results** *Karen Golden-Biddle*	• **Group Session Synthesizing the Key Messages (Structured Abstracts/ Executive Summaries)** *Marja Verhoef/ Ann Casebeer* **Project Work**	**Group Session Interpreting Results: Beyond Analysis** *Sheila Evans et al* **Mini-Clinics/ Consultations** *Creating Team et al* • Presenting Info & Data – Qualitative – Quantitative	**Critical Appraisal of Dissemination Strategies** *Rob Hayward*	**Project Presentations** (5 minute oral presentations) 8:30 a.m. to 12:30 p.m.

Time (unless otherwise stated)	**Sunday November 3**	**Monday November 4**	**Tuesday November 5**	**Wednesday November 6**	**Thursday November 7**	**Friday November 8**
Afternoon 1:00 pm – 4:30 pm	**Talking Circle** **Projects** (Individual Project Work/ Consultations) **Dissemination through Different Lens: A Panel Discussion** *Using Team et al*	**Dissemination: Developing a Plan for Sharing Results** (cont'd**)** **Using Theme Retrospective and Summing Up** *Trish Reay*	**Project Work** **Group Session Ethics Review Process Debrief** *Creating Team et al*	**Mini-Clinics/ Consultations** (cont'd**)** • Data Analysis Consults • Budgeting • Writing Papers and Reports • Privacy Impact Assessment • ? Subject to participant needs • Wrap-Up and • Next Steps *Creating Team et al*	**Information Management** *Rob Hayward* **Systems Integration Strategies** *Rob Hayward*	• **Talking Circle** • **Leave for Home**
Evening 7:00pm – 9:00pm	• **Projects** (Group Project Work/ Consultations)	**Faculty Meeting**	Free Evening	Free Evening	**Evening at Fort Edmonton Saloon**	

Appendix 2 Programme evaluation - metaevaluation quantitative checklist

PROGRAM EVALUATIONS METAEVALUATION CHECKLIST
(Based on The Program Evaluation Standards)

Daniel L. Stufflebeam
1999

This checklist is for performing final, summative metaevaluations. It is organized according to the Joint Committee Program Evaluation Standards. For each of the 30 standards the checklist includes 10 checkpoints drawn from the substance of the standard. It is suggested that each standard be scored on each checkpoint. Then judgments about the adequacy of the subject evaluation in meeting the standard can be made as follows: 0-2 Poor, 3-4 Fair, 5-6 Good, 7-8 Very Good, 9-10 Excellent. It is recommended that an evaluation be failed if it scores Poor on standards P1 Service Orientation, A5 Valid Information, A10 Justified Conclusions, or A11 Impartial Reporting. Users of this checklist are advised to consult the full text of The Joint Committee (1994) Program Evaluation Standards, Thousand Oaks, CA: Sage Publications.

TO MEET THE REQUIREMENTS FOR UTILITY, PROGRAM EVALUATIONS SHOULD:

U1 Stakeholder Identification

- Clearly identify the evaluation client
- Engage leadership figures to identify other stakeholders
- Consult potential stakeholders to identify their information needs
- Use stakeholders to identify other stakeholders
- With the client, rank stakeholders for relative importance
- Arrange to involve stakeholders throughout the evaluation
- Keep the evaluation open to serve newly identified stakeholders
- Address stakeholders' evaluation needs
- Serve an appropriate range of individual stakeholders
- Serve an appropriate range of stakeholder organizations

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

U2 Evaluator Credibility

- Engage competent evaluators
- Engage evaluators whom the stakeholders trust
- Engage evaluators who can address stakeholders= concerns
- Engage evaluators who are appropriately responsive to issues of gender, socioeconomic status, race, and language and cultural differences
- Assure that the evaluation plan responds to key stakeholders= concerns
- Help stakeholders understand the evaluation plan
- Give stakeholders information on the evaluation plan=s technical quality and practicality
- Attend appropriately to stakeholders= criticisms and suggestions
- Stay abreast of social and political forces
- Keep interested parties informed about the evaluation=s progress

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

U3 Information Scope and Selection

- Understand the client=s most important evaluation requirements
- Interview stakeholders to determine their different perspectives
- Assure that evaluator and client negotiate pertinent audiences, questions, and required information

- ☐ Assign priority to the most important stakeholders
- ☐ Assign priority to the most important questions
- ☐ Allow flexibility for adding questions during the evaluation
- ☐ Obtain sufficient information to address the stakeholders= most important evaluation questions
- ☐ Obtain sufficient information to assess the program=s merit
- ☐ Obtain sufficient information to assess the program=s worth
- ☐ Allocate the evaluation effort in accordance with the priorities assigned to the needed information

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

U4 Values Identification

- ☐ Consider alternative sources of values for interpreting evaluation findings
- ☐ Provide a clear, defensible basis for value judgments
- ☐ Determine the appropriate party(s) to make the valuational interpretations
- ☐ Identify pertinent societal needs
- ☐ Identify pertinent customer needs
- ☐ Reference pertinent laws
- ☐ Reference, as appropriate, the relevant institutional mission
- ☐ Reference the program=s goals
- ☐ Take into account the stakeholders= values
- ☐ As appropriate, present alternative interpretations based on conflicting but credible value bases

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

U5 Report Clarity

- ☐ Clearly report the essential information
- ☐ Issue brief, simple, and direct reports
- ☐ Focus reports on contracted questions
- ☐ Describe the program and its context
- ☐ Describe the evaluation=s purposes, procedures, and findings
- ☐ Support conclusions and recommendations
- ☐ Avoid reporting technical jargon
- ☐ Report in the language(s) of stakeholders
- ☐ Provide an executive summary
- ☐ Provide a technical report

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

U6 Report Timeliness and Dissemination

- ☐ Make timely interim reports to intended users
- ☐ Deliver the final report when it is needed
- ☐ Have timely exchanges with the program=s policy board
- ☐ Have timely exchanges with the program=s staff
- ☐ Have timely exchanges with the program=s customers
- ☐ Have timely exchanges with the public media
- ☐ Have timely exchanges with the full range of right-to-know audiences
- ☐ Employ effective media for reaching and informing the different audiences
- ☐ Keep the presentations appropriately brief
- ☐ Use examples to help audiences relate the findings to practical situations

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

U7 Evaluation Impact

- ☐ Involve stakeholders throughout the evaluation
- ☐ Encourage and support stakeholders= use of the findings
- ☐ Show stakeholders how they might use the findings in their work
- ☐ Forecast and address potential uses of findings
- ☐ Provide interim reports
- ☐ Make sure that reports are open, frank, and concrete
- ☐ Supplement written reports with ongoing oral communication
- ☐ Conduct feedback workshops to go over and apply findings
- ☐ Make arrangements to provide follow-up assistance in interpreting and applying the findings

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

Scoring the Evaluation for UTILITY	**Strength of the evaluation's provisions for UTILITY:**
Add the following:	☐ 26 (93%) to 28: **Excellent**
Number of Excellent ratings (0-7) _____ x 4 = _____	☐ 19 (68%) to 25: **Very Good**
Number of Very Good (0-7) _____ x 3 = _____	☐ 14 (50%) to 18: **Good**
Number of Good (0-7) _____ x 2 = _____	☐ 7 (25%) to 13: **Fair**
Number of Fair (0-7) _____ x 1 = _____	☐ 0 (0%) to 5: **Poor**
Total score: = _____	**__ (Total score) ÷28 = ___ x 100 = ____**

TO MEET THE REQUIREMENTS FOR FEASIBILITY, PROGRAM EVALUATIONS SHOULD:

F1 Practical Procedures

- ☐ Tailor methods and instruments to information requirements
- ☐ Minimize disruption
- ☐ Minimize the data burden
- ☐ Appoint competent staff
- ☐ Train staff
- ☐ Choose procedures that the staff are qualified to carry out
- ☐ Choose procedures in light of known constraints
- ☐ Make a realistic schedule
- ☐ Engage locals to help conduct the evaluation
- ☐ As appropriate, make evaluation procedures a part of routine events

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

F2 Political Viability

- ☐ Anticipate different positions of different interest groups
- ☐ Avert or counteract attempts to bias or misapply the findings
- ☐ Foster cooperation
- ☐ Involve stakeholders throughout the evaluation
- ☐ Agree on editorial and dissemination authority
- ☐ Issue interim reports
- ☐ Report divergent views
- ☐ Report to right-to-know audiences
- ☐ Employ a firm public contract
- ☐ Terminate any corrupted evaluation

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

F3 Cost Effectiveness

- ☐ Be efficient
- ☐ Make use of in-kind services
- ☐ Produce information worth the investment
- ☐ Inform decisions
- ☐ Foster program improvement
- ☐ Provide accountability information
- ☐ Generate new insights
- ☐ Help spread effective practices
- ☐ Minimize disruptions
- ☐ Minimize time demands on program personnel

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

Scoring the Evaluation for FEASIBILITY	**Strength of the evaluation's provisions for FEASIBILITY:**
Add the following:	☐ 11 (93%) to 28: **Excellent**
Number of Excellent ratings (0-3) _____ x 4 = _____	☐ 8 (68%) to 25: **Very Good**
Number of Very Good (0-3) _____ x 3 = _____	☐ 6 (50%) to 18: **Good**
Number of Good (0-3) _____ x 2 = _____	☐ 3 (25%) to 13: **Fair**
Number of Fair (0-3) _____ x 1 = _____	☐ 0 (0%) to 5: **Poor**
Total score: = _____	**__ (Total score) ÷12 = ___ x 100 = ____**

TO MEET THE REQUIREMENTS FOR PROPRIETY, PROGRAM EVALUATIONS SHOULD:

P1 Service Orientation

- ☐ Assess needs of the program=s customers
- ☐ Assess program outcomes against targeted customers= assessed needs
- ☐ Help assure that the full range of rightful program beneficiaries are served
- ☐ Promote excellent service
- ☐ Make the evaluation=s service orientation clear to stakeholders
- ☐ Identify program strengths to build on
- ☐ Identify program weaknesses to correct
- ☐ Give interim feedback for program improvement
- ☐ Expose harmful practices
- ☐ Inform all right-to-know audiences of the program=s positive and negative outcomes

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

P2 Formal Agreements, reach advance written agreements on:

- ☐ Evaluation purpose and questions
- ☐ Audiences
- ☐ Evaluation reports
- ☐ Editing
- ☐ Release of reports
- ☐ Evaluation procedures and schedule
- ☐ Confidentiality/anonymity of data
- ☐ Evaluation staff
- ☐ Metaevaluation
- ☐ Evaluation resources

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

P3 Rights of Human Subjects

- Make clear to stakeholders that the evaluation will respect and protect the rights of human subjects
- Clarify intended uses of the evaluation
- Keep stakeholders informed
- Follow due process
- Uphold civil rights
- Understand participant values
- Respect diversity
- Follow protocol
- Honor confidentiality/anonymity agreements
- Do no harm

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

P4 Human Interactions

- Consistently relate to all stakeholders in a professional manner
- Maintain effective communication with stakeholders
- Follow the institution=s protocol
- Minimize disruption
- Honor participants= privacy rights
- Honor time commitments
- Be alert to and address participants= concerns about the evaluation
- Be sensitive to participants= diversity of values and cultural differences
- Be even-handed in addressing different stakeholders
- Do not ignore or help cover up any participant=s incompetence, unethical behavior, fraud, waste, or abuse

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

P5 Complete and Fair Assessment

- Assess and report the program=s strengths
- Assess and report the program=s weaknesses
- Report on intended outcomes
- Report on unintended outcomes
- Give a thorough account of the evaluation=s process
- As appropriate, show how the program=s strengths could be used to overcome its weaknesses
- Have the draft report reviewed
- Appropriately address criticisms of the draft report
- Acknowledge the final report=s limitations
- Estimate and report the effects of the evaluation=s limitations on the overall judgment of the program

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

P6 Disclosure of Findings

- Define the right-to-know audiences
- Establish a contractual basis for complying with right-to-know requirements
- Inform the audiences of the evaluation=s purposes and projected reports
- Report all findings in writing
- Report relevant points of view of both supporters and critics of the program
- Report balanced, informed conclusions and recommendations
- Show the basis for the conclusions and recommendations

- Disclose the evaluation=s limitations
- In reporting, adhere strictly to a code of directness, openness, and completeness
- Assure that reports reach their audiences

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

P7 Conflict of Interest

- Identify potential conflicts of interest early in the evaluation
- Provide written, contractual safeguards against identified conflicts of interest
- Engage multiple evaluators
- Maintain evaluation records for independent review
- As appropriate, engage independent parties to assess the evaluation for its susceptibility or corruption by conflicts of interest
- When appropriate, release evaluation procedures, data, and reports for public review
- Contract with the funding authority rather than the funded program
- Have internal evaluators report directly to the chief executive officer
- Report equitably to all right-to-know audiences
- Engage uniquely qualified persons to participate in the evaluation, even if they have a potential conflict of interest; but take steps to counteract the conflict

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

P8 Fiscal Responsibility

- Specify and budget for expense items in advance
- Keep the budget sufficiently flexible to permit appropriate reallocations to strengthen the evaluation
- Obtain appropriate approval for needed budgetary modifications
- Assign responsibility for managing the evaluation finances
- Maintain accurate records of sources of funding and expenditures
- Maintain adequate personnel records concerning job allocations and time spent on the job
- Employ comparison shopping for evaluation materials
- Employ comparison contract bidding
- Be frugal in expending evaluation resources
- As appropriate, include an expenditure summary as part of the public evaluation report

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

Scoring the Evaluation for PROPRIETY **Add the following:**	**Strength of the evaluation's provisions for PROPRIETY:**
Number of Excellent ratings (0-8) _____ x 4 = _____	☐ 30 (93%) to 28: **Excellent**
Number of Very Good (0-8) _____ x 3 = _____	☐ 22 (68%) to 25: **Very Good**
Number of Good (0-8) _____ x 2 = _____	☐ 16 (50%) to 18: **Good**
Number of Fair (0-8) _____ x 1 = _____	☐ 8 (25%) to 13: **Fair**
Total score: = _____	☐ 0 (0%) to 5: **Poor**
	__ (Total score) ÷32 = _____ x 100 = ____

TO MEET THE REQUIREMENTS FOR ACCURACY, PROGRAM EVALUATIONS SHOULD:

A1 Program Documentation

- Collect descriptions of the intended program from various written sources
- Collect descriptions of the intended program from the client and various stakeholders
- Describe how the program was intended to function
- Maintain records from various sources of how the program operated
- As feasible, engage independent observers to describe the program=s actual operations
- Describe how the program actually functioned
- Analyze discrepancies between the various descriptions of how the program was intended to

function

- Analyze discrepancies between how the program was intended to operate and how it actually operated
- Ask the client and various stakeholders to assess the accuracy of recorded descriptions of both the intended and the actual program
- Produce a technical report that documents the program=s operations

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

A2 Context Analysis

- Use multiple sources of information to describe the program=s context
- Describe the context=s technical, social, political, organizational, and economic features
- Maintain a log of unusual circumstances
- Record instances in which individuals or groups intentionally or otherwise interfered with the program
- Record instances in which individuals or groups intentionally or otherwise gave special assistance to the program
- Analyze how the program=s context is similar to or different from contexts where the program might be adopted
- Report those contextual influences that appeared to significantly influence the program and that might be of interest to potential adopters
- Estimate effects of context on program outcomes
- Identify and describe any critical competitors to this program that functioned at the same time and in the program=s environment
- Describe how people in the program=s general area perceived the program=s existence, importance, and quality

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

A3 Described Purposes and Procedures

- At the evaluation=s outset, record the client=s purposes for the evaluation
- Monitor and describe stakeholders= intended uses of evaluation findings
- Monitor and describe how the evaluation=s purposes stay the same or change over time
- Identify and assess points of agreement and disagreement among stakeholders regarding the evaluation=s purposes
- As appropriate, update evaluation procedures to accommodate changes in the evaluations purposes
- Record the actual evaluation procedures, as implemented
- When interpreting findings, take into account the different stakeholders= intended uses of the evaluation
- When interpreting findings, take into account the extent to which the intended procedures were effectively executed
- Describe the evaluation=s purposes and procedures in the summary and full-length evaluation reports
- As feasible, engage independent evaluators to monitor and evaluate the evaluations purposes and procedures

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

A4 Defensible Information Sources

- Obtain information from a variety of sources
- Use pertinent, previously collected information once validated
- As appropriate, employ a variety of data collection methods
- Document and report information sources
- Document, justify, and report the criteria and methods used to select information sources

- ☐ For each source, define the population
- ☐ For each population, as appropriate, define any employed sample
- ☐ Document, justify, and report the means used to obtain information from each source
- ☐ Include data collection instruments in a technical appendix to the evaluation report
- ☐ Document and report any biasing features in the obtained information

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

A5 Valid Information

- ☐ Focus the evaluation on key questions
- ☐ As appropriate, employ multiple measures to address each question
- ☐ Provide a detailed description of the constructs and behaviors about which information will be acquired
- ☐ Assess and report what type of information each employed procedure acquires
- ☐ Train and calibrate the data collectors
- ☐ Document and report the data collection conditions and process
- ☐ Document how information from each procedure was scored, analyzed, and interpreted
- ☐ Report and justify inferences singly and in combination
- ☐ Assess and report the comprehensiveness of the information provided by the procedures as a set in relation to the information needed to answer the set of evaluation questions
- ☐ Establish meaningful categories of information by identifying regular and recurrent themes in information collected using qualitative assessment procedures

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

A6 Reliable Information

- ☐ Identify and justify the type(s) and extent of reliability claimed
- ☐ For each employed data collection device, specify the unit of analysis
- ☐ As feasible, choose measuring devices that in the past have shown acceptable levels of reliability for their intended uses
- ☐ In reporting reliability of an instrument, assess and report the factors that influenced the reliability, including the characteristics of the examinees, the data collection conditions, and the evaluators biases
- ☐ Check and report the consistency of scoring, categorization, and coding
- ☐ Train and calibrate scorers and analysts to produce consistent results
- ☐ Pilot test new instruments in order to identify and control sources of error
- ☐ As appropriate, engage and check the consistency between multiple observers
- ☐ Acknowledge reliability problems in the final report
- ☐ Estimate and report the effects of unreliability in the data on the overall judgment of the program

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

A7 Systematic Information

- ☐ Establish protocols for quality control of the evaluation information
- ☐ Train the evaluation staff to adhere to the data protocols
- ☐ Systematically check the accuracy of scoring and coding
- ☐ When feasible, use multiple evaluators and check the consistency of their work
- ☐ Verify data entry
- ☐ Proofread and verify data tables generated from computer output or other means
- ☐ Systematize and control storage of the evaluation information
- ☐ Define who will have access to the evaluation information
- ☐ Strictly control access to the evaluation information according to established protocols
- ☐ Have data providers verify the data they submitted

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor
A8 Analysis of Quantitative Information
☐ Begin by conducting preliminary exploratory analyses to assure the data=s correctness and to gain a greater understanding of the data ☐ Choose procedures appropriate for the evaluation questions and nature of the data ☐ For each procedure specify how its key assumptions are being met ☐ Report limitations of each analytic procedure, including failure to meet assumptions ☐ Employ multiple analytic procedures to check on consistency and replicability of findings ☐ Examine variability as well as central tendencies ☐ Identify and examine outliers and verify their correctness ☐ Identify and analyze statistical interactions ☐ Assess statistical significance and practical significance ☐ Use visual displays to clarify the presentation and interpretation of statistical results
☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor
A9 Analysis of Qualitative Information
☐ Focus on key questions ☐ Define the boundaries of information to be used ☐ Obtain information keyed to the important evaluation questions ☐ Verify the accuracy of findings by obtaining confirmatory evidence from multiple sources, including stakeholders ☐ Choose analytic procedures and methods of summarization that are appropriate to the evaluation questions and employed qualitative information ☐ Derive a set of categories that is sufficient to document, illuminate, and respond to the evaluation questions ☐ Test the derived categories for reliability and validity ☐ Classify the obtained information into the validated analysis categories ☐ Derive conclusions and recommendations and demonstrate their meaningfulness ☐ Report limitations of the referenced information, analyses, and inferences
☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor
A10 Justified Conclusions
☐ Focus conclusions directly on the evaluation questions ☐ Accurately reflect the evaluation procedures and findings ☐ Limit conclusions to the applicable time periods, contexts, purposes, and activities ☐ Cite the information that supports each conclusion ☐ Identify and report the program=s side effects ☐ Report plausible alternative explanations of the findings ☐ Explain why rival explanations were rejected ☐ Warn against making common misinterpretations ☐ Obtain and address the results of a prerelease review of the draft evaluation report ☐ Report the evaluation=s limitations
☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor
A11 Impartial Reporting
☐ Engage the client to determine steps to ensure fair, impartial reports ☐ Establish appropriate editorial authority ☐ Determine right-to-know audiences ☐ Establish and follow appropriate plans for releasing findings to all right-to-know audiences ☐ Safeguard reports from deliberate or inadvertent distortions ☐ Report perspectives of all stakeholder groups

- ☐ Report alternative plausible conclusions
- ☐ Obtain outside audits of reports
- ☐ Describe steps taken to control bias
- ☐ Participate in public presentations of the findings to help guard against and correct distortions by other interested parties

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

A12 Metaevaluation

- ☐ Designate or define the standards to be used in judging the evaluation
- ☐ Assign someone responsibility for documenting and assessing the evaluation process and products
- ☐ Employ both formative and summative metaevaluation
- ☐ Budget appropriately and sufficiently for conducting the metaevaluation
- ☐ Record the full range of information needed to judge the evaluation against the stipulated standards
- ☐ As feasible, contract for an independent metaevaluation
- ☐ Determine and record which audiences will receive the metaevaluation report
- ☐ Evaluate the instrumentation, data collection, data handling, coding, and analysis against the relevant standards
- ☐ Evaluate the evaluation=s involvement of and communication of findings to stakeholders against the relevant standards
- ☐ Maintain a record of all metaevaluation steps, information, and analyses

☐ 9-10 Excellent ☐ -8 Very Good ☐ 5-6 Good ☐ 3-4 Fair ☐ 0-2 Poor

Scoring the Evaluation for ACCURACY	**Strength of the evaluation's provisions for ACCURACY:**
Add the following:	
Number of Excellent ratings (0-12) _____ x 4 = _____	☐ 45 (93%) to 28: **Excellent**
Number of Very Good (0-12) _____ x 3 = _____	☐ 33 (68%) to 25: **Very Good**
Number of Good (0-12) _____ x 2 = _____	☐ 24 (50%) to 18: **Good**
Number of Fair (0-12) _____ x 1 = _____	☐ 12(25%) to 13: **Fair**
Total score: = _____	☐ 0 (0%) to 5: **Poor**
	__ (Total score) ÷32 = _____ x 100 = ____

Appendix 3 Table comparing evolution of evaluation standards 1984-2011 (Yarbrough et al., 2011)

	1981		1994		2011
Utility Standards	The Utility Standards are intended to ensure that an evaluation will serve the practical information needs of given audiences		The Utility Standards are intended to ensure that an evaluation will serve the information needs of intended users		Their goal is to increase the likelihood that the evaluation will have positive consequences and substantial influence, as needs and opportunities appear over the course of the evaluation (pg 8)
A1	Audience Identification Audiences involved in or affected by the evaluation should be identified, so that their needs can be addressed	U1	Stakeholder Identification Persons involved in or affected by the evaluation should be identified so that their need can be addressed	U1	Evaluator Credibility Evaluations should be conducted by qualified people who establish and maintain credibility in the evaluation context
A2	Evaluator Credibility The persons conducting the evaluation should be both trustworthy and competent to perform the evaluation, so that their findings achieve maximum credibility and acceptance	U2	Evaluator Credibility The persons conducing the evaluation should be both trustworthy and competent to perform the evaluation, so that the evaluation findings achieve maximum credibility and acceptance	U2	Attention to Stakeholders Evaluations should devote attention to the full range of individuals and groups invested in the program and affected by its evaluation
A3	Information Scope and Selection Information collected should be of such scope and selected in such ways as to address pertinent questions about the object of the evaluation and be responsive to the needs and interests of specified audiences	U3	Information Scope and Selection Information collected should be broadly selected to address pertinent question about the program and be responsive to the needs and interests of clients and other specified stakeholders	U3	Negotiated Purpose Evaluation purposes should be identified and continually negotiated based on the needs of the stakeholders
A4	Valuation Interpretation The perspectives, procedures, and rationale used to interpret the findings should be carefully described, so that the bases for value judgements are clear	U4	Values Identification The perspectives, procedures, and rationale used to interpret the findings should be carefully described, so that the bases for value judgments are clear	U4	Explicit Values Evaluation should clarify and specify the individual and cultural values underpinning purposes, processes, and judgement

	1981		1994		2011
A5	Report Clarity The evaluation report should describe the object being evaluated and its context, and the purposes, procedures, and findings of the evaluation, so that the audiences will readily understand what was done, why it was done, what information was obtained, what conclusions were drawn and what recommendations were made	U5	Report Clarity Evaluation reports should clearly describe the program being evaluated, including its context, and the purposes, procedures, and findings of the evaluation, so that essential information is provided and easily understood	U5	Relevant Information Evaluation information should serve the identified and emergent needs of stakeholders
A6	Report Dissemination Evaluation findings should b disseminated to clients and other right-to-know audiences, so that they can asses and use the findings			U6	Meaningful Processes and Products Evaluation should construct activities, descriptions, and judgements in ways that encourage participants to rediscover, reinterpret, or revise their understanding and behaviours
A7	Report Timeliness Release of reports should be timely, so that audiences can best use the reported information	U6	Report Timeliness and Dissemination Significant interim findings and evaluation reports should be disseminated to intended users, so that they can be used in a timely fashion	U7	Timely and Appropriate Communicating and Reporting Evaluations should attend to the continuing information needs of their multiple audiences
A8	Evaluation Impact Evaluations should be planned and conducted in ways that encourage follow-through by members of the audiences	U7	Evaluation Impact Evaluations should be planned, conducted and reported in ways that encourage follow-through by stakeholders, so that the likelihood that the evaluation will be used is increased	U8	Concern for Consequences and Influence Evaluation should promote responsible and adaptive use while guarding against unintended negative consequences and misuse

	1981		1994		2011
Feasibility Standards	The Feasibility Standards are intended to ensure that an evaluation will be realistic, prudent, diplomatic, and frugal		The Feasibility Standards are intended to ensure that the evaluation will be realistic, prudent, diplomatic, and fugal		Attention to feasibility highlights the logistical and administrative requirements of evaluations that must be managed (pg72)
B1	Practical Procedures The evaluation procedures should be practical, so that disruption is kept to a minimum, and that needed information can be obtained	F1	Practical Procedures The evaluation procedures should be practical, to keep disruption to a minimum while needed information is obtained	F1	Project Management Evaluation should use effective project management strategies
B2	Political Viability The evaluation should be planned and conducted with anticipation of the different positions of various interest groups, so that their cooperation may be obtained , and so that possible attempts by any of these groups to curtail evaluation operations or to bias or misapply the results can be averted or counteracted	F2	Political Viability The evaluation should be planned and conducted with anticipation of different positions of various interest groups, so that their cooperation may be obtained, and so tha that possible attempts by any of these groups to curtail evaluation operations or to bias or misapply the results can be averted or counteracted	F2	Practical Procedures Evaluation procedures should be practical and responsive to the way the progam operates
B3	Cost Effectiveness The evaluation should produce information of sufficient value to justify the resources expended	F3	Cost Effectiveness The evaluation should be efficient and produce information of sufficient value, so that the resources expended can be justified	F3	Contextual Viability Evaluations should recognize, monitor, and balance the cultural and political interests and needs of individuals and groups
				F4	Resource Use Evaluations should use resources effectively and efficiently

	1981		1994		2011
Propriety Standards	The Propriety Standards are intended to ensure that an evaluation will be conducted legally, ethically, and with due regard for the welfare of those involved in the evaluation, as well as those affected by its results		The Propriety Standards are intended to ensure that an evaluation will be conducted legally, ethically and with due regard for the welfare of those involved in the evaluation, as well as those affected by its results		Propriety refers to what is proper, fair, legal, right, acceptable, and just in evaluation (pg 106)
C1	Formal Obligation Obligations of the formal parties to an evaluation (what is to be done, how, by whom, when) should be agreed to in writing, so that these parties are obligated to adhere to all conditions of the agreement or formally to renegotiate it	P1	Service Orientation Evaluations should be designed to assist organizations to address and effectively serve the needs of the full range of targeted participants	P1	Responsive and Inclusive Orientation Evaluation should be responsive to stakeholders and their communities
C2	Conflict of Interest Conflict of interest, frequently unavoidable, should be dealt with openly and honestly, so that it does not compromise the evaluation processes and results	P2	Formal Agreements Obligations of the formal parties to an evaluation (what is to be done, how, by whom, when) should be agreed to in writing, so that these parties are obligated to adhere to all conditions of the agreement or formally to renegotiate it	P2	Formal Agreements Evaluation agreements should be negotiated to make obligations explicitly and take into account the needs, expectations, and cultural contexts of clients and other stakeholder
C3	Full and Frank Discloser Oral and written evaluation reports should be open, direct, and honest in their disclosure of pertinent findings, including the limitations of the evaluation	P3	Rights of Human Subjects Evaluations should be designed and conducted to respect and protect the rights and welfare of human subjects	P3	Human Rights and Respect Evaluations should be designed and conducted to protect human and legal rights and maintain the dignity of participants and other stakeholders
C4	Public's Right to Know The formal parties to an evaluation should respect and assure the public's right to know, within the limits of other related principles and statutes, such as those dealing with public safety and the right to privacy	P4	Human Interactions Evaluators should respect human dignity and worth in their interactions with other persons associated with an evaluation so that participants are not threatened or harmed		

	1981		1994		2011
C5	Rights of Human Subjects Evaluations should be designed and conducted, so that the rights and welfare of the human subjects are respected and protected	P5	Complete and Fair Assessment The evaluation should b complete and fair in its examination and recording of strengths and weaknesses of the program being evaluated, so that strengths can be build upon and problem areas addressed	P4	Clarity and Fairness Evaluations should be understandable and fair in addressing stakeholder needs and purposes
	Human Interactions Evaluators should respect human dignity and worth in their interactions with other persons associated with an evaluation	P6	Disclosure of Findings The formal parties to an evaluation should ensure that the full set of evaluation findings along with pertinent limitations are made accessible to the persons affected by the evaluation, and any others with expressed legal rights to receive the results	P5	Transparency and Disclosure Evaluations should provide complete descriptions of findings, limitations, and conclusions to all stakeholder, unless doing so would violate legal and propriety obligations
C7	Balanced Reporting The evaluation should be complete and fair in its presentation of strengths and weaknesses of the object under investigation, so that strengths can be build upon and problem areas addressed	P7	Conflict of Interest Conflict of interest should be dealt with openly and honestly, so that it does not compromise the evaluation processes and results	P6	Conflicts of Interest Evaluation should openly and honestly identify and address real or perceived conflicts of interests that may compromise the evaluation
C8	Fiscal Responsibility The evaluator's allocation and expenditure of resources should reflect sound accountability procedures and otherwise be prudent and ethically responsible	P8	Fiscal Responsibility The evaluator's allocation and expenditure of resources should reflect sound accountability procedures and otherwise be prudent and ethically responsible, so that expenditures are accounted for and appropriate	P7	Fiscal Responsibility Evaluations should account for all expended resources and comply with sound fiscal procedures and processes
Accuracy Standards	The Accuracy Standards are intended to ensure that an evaluation will reveal and convey technically adequate information about the features of the object being studied that determine its worth or merit		The Accuracy Standards are intended to ensure that an evaluation will reveal and convey technically adequate information about features that determine worth or merit of the program being evaluated		Accuracy is the truthfulness of evaluation representations, propositions, and findings, especially those that support judgments about the quality of programs or program components (pg 158)

	1981		1994		2011
D1	Object Identification The object of the evaluation (program, project, material) should be sufficiently examined, so that the form(s) of the object being considered in the evaluation can be clearly identified	A1	Program Documentation The program being evaluated should be described and documented clearly and accurately, so that the program is clearly identified	A1	Justified Conclusions and Decisions Evaluation conclusions and decisions should be explicitly justified in the cultures and contexts where they have consequences
D2	Context Analysis The context in which the program, project or material exists should be examined in enough detail, so that its likely influences on the object can be identified	A2	Context Analysis The context in which the program exists should be examined in enough detail, so that its likely influences on the program can be identified	A2	Valid Information Evaluation information should serve the intended purposes and support valid interpretations
D3	Described Purposes and Procedures The purposes and procedures of the evaluation should be monitored and described in enough detail, so that they can be identified and assessed	A3	Described Purpose and Procedures The purposes and procedures of the evaluation should be monitored and described in enough detail, so that they can be identified and assesse.	A3	Reliable Information Evaluation procedures should yield sufficiently dependable and consistent information for the intended uses
D4	Defensible Information Sources The sources of information should be described in enough detail, to that the adequacy of the information can be assessed	A4	Defensible Information Sources The sources of information used in a program evaluation should be described in enough detail, so that the adequacy of the information can be assessed	A4	Explicit Program and Context descriptions Evaluations should document programs and their contexts with appropriate detail and scope for the evaluation purposes
D5	Valid Measurement The information-gathering instruments and procedures should be chosen or developed and then implemented in ways that will assure that the interpretation arrived at is valid for the given use	A5	Valid Information The information gathering procedures should be chosen or developed and then implemented so that they will assure that the interpretation arrived at is valid for the intended use	A5	Information Management Evaluations should employ systematic information collection, review, verification, and storage details

	1981		1994		2011
D6	Reliable Measurement The information-gathering instruments and procedures should be chosen or developed and then implemented in ways that will assure that the information obtained is sufficiently reliable for the intended use	A6	Reliable Information The information gathering procedures should be chosen or developed and then implemented so that they will assure that the information obtained is sufficiently reliable for the intended use	A6	Sound designs and Analysis Evaluations should employ technically adequate designes and analyses that are appropriate for the evaluation puposes
D7	Systematic Data Control The data collected, processed, and reported in an evaluation should be reviewed and corrected, so that the results of the evaluation will not be flawed	A7	Systematic Information The information collected, processed, and reported in an evaluation should be systematically reviewed and any erros found should be corrected	A7	Explicit Evaluation Reasoning Evaluation reasoning leading from information and analyses to finding, interpretations, conclusions, and judgements should be clearly and completely documented
D8	Analysis of Quantitative Information Quantitative information in an evaluation should be appropriately and systematically analyzed to ensure supportable interpretations	A8	Analysis of Quantitative Information Quantitative information in an evaluation should be appropriately and systematically analyzed so that evaluations questions are effectively answered	A8	Communications and Reporting Evaluation communications should have adequate scope and guard against misconceptions, biases, distortions, and errors
D9	Analysis of Qualitative Information Qualitative information in an evaluation should be appropriately and systematically analyzed to ensure supportable interpretations	A9	Analysis of Qualitative Information Qualitative information in an evaluation should be appropriately and systematically analyzed so that evaluations questions are effectively answered		
D10	Justified Conclusions The conclusions reached in an evaluation should be explicitly justified, so that the audiences can assess them	A10	Justified Conclusions The conclusions reached in an evaluation should be explicitly justified, so that stakeholders can assess them		

	1981		1994		2011
D11	Objective reporting The evaluation procedures should provide safeguards to protect the evaluation finding and reports against distortion by the personal feelings and biases of any party to the evaluation	A11	Impartial Reporting Reporting procedures should guard against distortion caused by personal feeling and biases of any party to the evaluation, so that evaluation reports fairly reflect the evaluation findings		
		A12	Meta evaluation The evaluation itself should be formatively and summatively evaluated against these and other pertinent standards, so that its conduct is appropriately guided and, and on completion, stakeholders can closely examine its strengths and weaknesses		
	Evaluation Accountability Standard – added in 2011			E1	Evaluation documentation Evaluations should fully document their negotiated purposes and implemented designs, procedures, data and outcome
				E2	Internal Metaevaluation Evaluators should use these and other applicable standards to examine the accountability of the evaluation design, procedures employed, information collected, and outcomes
				E3	External Metaevaluation Program evaluation sponsors, clients, evaluators, and other stakeholders should encourage the conduct of external metaevaluation using these and other applicable standards

Appendix 4 Quantitative analysis summary

Utility data

The utility standard relates to the ability of the evaluation to meet the information needs of the intended users and is made up of seven checklist items (Table 27).

The first item relates to stakeholder identification. The evaluations scored positively in the areas of clearly identifying clients, and engaging them to identify stakeholders, but poorly on the remainder of the items which related to engaging and ranking the stakeholders and attempting to meet their needs of the various needs.

Item two related to the evaluators' credibility. Given the limited number of evaluators there was a limited range of scores in this area. The evaluations scored high on the items related to competency and also on provision of the evaluation plan to stakeholders. They scored poorly on issues related to their responsiveness to stakeholder needs and on flexibility. Given the collaborative nature of the relationship between the evaluators and the SEARCH Program faculty it is possible that this was negotiated but not reported.

The overall scores on the third item, which relates to information gathering and scope were somewhat better than those for the first two items. All evaluations demonstrated that the evaluators understood the SEARCH Program requirements and a majority also demonstrated the merit and worth of the programme. Scores were low in the areas of negotiation of priority and in demonstration of flexibility.

Item four relates to value identification. The evaluations rated well in only two categories, those related to referencing of the institutional mission and programme goals that were outlined in all evaluations. The evaluations did not demonstrate any methods that might have been employed to consider other sources of data or stakeholder values.

All evaluations rated highly with regard to the fifth item, which measured report clarity. Two reports scored 6/10; the remainder scored 9 or 10. The reports were professional in presentation, well organised and clearly written.

Item six refers to report timeliness and dissemination. The evaluations scored highly in only one category, which was the appropriate delivery of the final report. Other categories included communication exchanges with programme staff and/or with the public or with the media. It might be expected that there was ongoing communication with the programme staff, but there were limited indications in the reports that this took place. In

relation to public dissemination, there is no indication that it was the responsibility of the evaluators, so the low rating is not a surprise.

The final item addresses issues regarding evaluation impact. As with external dissemination there is no indication that the evaluators had accepted the responsibility for evaluation impact and therefore as would be expected the scores in this area were very low. Where evaluations did score points were in providing interim reports and making sure that reports were open, frank and concrete.

The high rating for report clarity item meant that a number of evaluations had at least one excellent mark in their overall score. However, in general the ratings were only fair thus producing strength scores that ranged from 3 to 12. The individual strength of the evaluations' provision for utility ranged between 11% and 43%. The Managers' Survey and the Collaborative Network Survey had the lowest scores in this category – a situation that is the same in the other categories as well.

Feasibility data

This is the shortest of the four categories, with only three items and relates to whether the evaluation is realistic, prudent, diplomatic, and frugal (Table 28). The first item related to the practical procedures used in the evaluation where overall the SEARCH evaluations scored between 4/10 and 5/10. All evaluations scored well in the areas of tailoring methods, minimising disruption and data burden. The majority demonstrated that they had set reasonable schedules. None provide information about the training or qualification of the staff conducting the evaluation or the engagement of locals to collect data.

The second item relates to anticipating the various positions and interests of the group involved to elicit co-operation and identify potential bias. Overall the evaluations scored poorly in this area. Like many other aspects of the metaevaluation it is not clear whether this is because the issues were not addressed in the evaluation or simply not reported.

The third item relates to cost effectiveness, and whether the evaluation was efficient in producing valuable information that justified the expenditure. The evaluations scored reasonably on only two of the factors in this area: efficiency in the conduct of the evaluation, and the use of in-kind services (defined in this case as services within the existing programme structure).

In terms of feasibility there were no excellent or very good scores and the strength scores were between 2 and 3. In terms of overall results the percentages ranged from 16.7% to 25% indicating a very low strength of the evaluations.

Table 27 Utility totals

Evaluation	U1 total	U2 total	U3 total	U4 total	U5 total	U6 total	U7 total	U N excellent	U N very good	U N good	U N fair	U N poor	U excellent score	U very good score	U good score	U fair score	U strength score	U result
SEARCH I	4	4	7	3	9	4	3	1	1	0	5	0	4	3	0	5	12	42.9
SEARCH II	5	4	4	3	9	2	4	1	0	1	4	1	4	0	2	4	10	35.7
SEARCH I and II	3	3	5	2	9	1	1	1	0	1	2	3	4	0	2	2	8	28.6
SEARCH III	2	3	4	2	9	3	2	1	0	0	3	2	4	0	0	3	7	25
SEARCH IV	3	2	8	1	10	3	2	1	1	0	2	3	4	3	0	2	9	32.1
Faculty Impact	2	5	4	3	10	1	1	1	0	1	2	3	4	0	2	2	8	28.6
Project Tracking	4	2	4	3	9	1	1	1	0	0	3	3	4	0	0	3	7	25
Organisational Impact 1	2	3	4	2	9	1	1	1	0	0	2	4	4	0	0	2	6	21.4
Managers' Survey	2	3	1	3	6	0	0	0	0	1	2	4	0	0	2	2	4	14.3
Collaborative Network Evaluation	1	3	1	2	6	0	0	0	0	1	1	5	0	0	2	1	3	10.7

Table 28 Feasibility totals

Evaluation	F1 total	F2 total	F 3 total	F N excellent	F N very good	F N good	F N fair	F N poor	F excellent score	F very good score	F good score	F fair score	F strength score	F result
SEARCH I	5	3	1	0	0	1	1	1	0	0	2	1	3	25
SEARCH II	5	2	3	0	0	1	1	1	0	0	2	1	3	25
SEARCH I and II	5	1	2	0	0	1	0	2	0	0	2	0	2	16.7
SEARCH III	3	2	3	0	0	0	2	1	0	0	0	2	2	16.7
SEARCH IV	5	2	4	0	0	1	1	1	0	0	2	1	3	25
Faculty Impact	5	0	1	0	0	1	0	2	0	0	2	0	2	16.7
Project Tracking	5	0	1	0	0	1	0	2	0	0	2	0	2	16.7
Organisational Impact 1	5	1	1	0	0	1	0	2	0	0	2	0	2	16.7
Managers' Survey	5	0	2	0	0	1	0	2	0	0	2	0	2	16.7
Collaborative Network Evaluation	4	0	4	0	0	0	2	1	0	0	0	2	2	16.7

Propriety data

There are eight items in the Propriety standard, which assess the legal and ethical issues related to the evaluation, by examining whether there was due regard for those involved in the evaluation or affected by its results. Results are presented in Table 29

Item one relates to the service orientation of the evaluators. The evaluations scored well in the areas of identifying the programme's strengths and weaknesses. They scored poorly in the areas of assessing customer needs and ensuring that appropriate audiences received results.

The second item relates to formal agreements that guide the evaluation, including all aspects of the conduct of the work. The majority of evaluations scored on only one factor - defining the evaluation purpose and question. Examination of the SEARCH Program records identified written contracts for the majority of evaluations, but these were focused on payment and did not include explicit mention of the other factors included in this item. It is worth noting that during the qualitative data analysis reported later an action item for the programme director was to define standard evaluator requirements for all evaluation contracts.

The third item relates to respect for the rights of human subjects. The evaluations scored in the areas of following the evaluation protocol and process. Only one evaluation reported vetting through an ethics committee.

Factor four is linked to this and relates to how the evaluators interact during the evaluation. The evaluations rated well in the areas of adapting a professional manner and following a set protocol, but the reports did not provide any information regarding diversity or privacy. Although no evaluations specifically identified individuals or health regions in their reports, a number of reports contained information that could have allowed the reader to identify the source of the data. In addition it was not possible to provide anonymity within the faculty evaluation, as the data referred back to the specific institutions involved.

The fifth item refers to the content of the report, where the evaluations scored well in the areas of identifying strengths and weaknesses and thorough reporting. They scored poorly in the areas of draft report editing and identifying limitations of the report.

Item six relates to disclosure of findings, where overall the evaluations scored poorly with the exception of providing written reports. The next item relates to the identification of conflict of interests, and none of the evaluations addressed these issues. The final item deals with fiscal responsibility; the only evidence of this was found in contracts or letters of agreement related to the evaluations.

Where data related to contractual obligations and economics were not available in the report, the data were taken from correspondence records in the electronic archive. Although it was not possible to identify specific contractual arrangements for all evaluation contracts the existing policies relating to accounting practice mean that each external contract would have been managed according to standard accounting practices(SEARCH Canada, 2008). Numerous evaluation project proposals were identified in the electronic files along with written responses from SEARCH executive officers. Therefore where no specific data were available each report was credited with a 0.

Overall in this category there were no scores of excellent or very good and the overall percentages ranged from 3% to 22% indicating a very low strength of the evaluations' provision for propriety.

Table 29 Propriety totals

Evaluation	P1 total	P2 total	P3 total	P4 total	P5 total	P6 total	P7 total	P8 total	P N excellent	P N very good	P N good	P N fair	P N poor	P excellent score	P very good score	P good score	P fair score	P strength score	P result
SEARCH I	4	3	2	2	7	4	0	1	0	1	0	3	4	0	3	0	3	6	18.8
SEARCH II	4	2	4	3	6	6	0	1	0	0	2	3	3	0	0	4	3	7	21.9
SEARCH I and II	3	1	3	2	5	5	0	1	0	0	2	2	4	0	0	4	2	6	18.8
SEARCH III	4	1	1	2	5	1	1	1	0	0	1	1	6	0	0	2	1	3	9.38
SEARCH IV	3	1	3	2	4	3	0	1	0	0	0	4	4	0	0	0	4	4	12.5
Faculty Impact	4	1	3	3	4	4	0	1	0	0	0	5	3	0	0	0	5	5	15.6
Project Tracking	2	1	2	3	5	3	0	1	0	0	1	2	5	0	0	2	2	4	12.5
Organisational Impact 1	3	1	3	2	5	1	0	1	0	0	1	2	5	0	0	2	2	4	12.5
Managers' Survey	1	1	3	3	0	0	0	0	0	0	0	2	6	0	0	0	2	2	6.25
Collaborative Network Evaluation	0	1	2	3	1	1	0	0	0	0	0	1	7	0	0	0	1	1	3.13

Table 30 Accuracy totals

Evaluation	A1 total	A2 total	A3 total	A4 total	A5 total	A6 total	A7 total	A8 total	A9 total	A10 total	A11 total	A12 total	A N excellent	A N very good	A N good	A N fair	A N poor	A excellent score	A very good score	A good score	A fair score	A strength score	A result
SEARCH I	7	3	4	7	3	0	0	0	4	3	0	0	0	2	0	5	5	0	6	0	5	11	22.92
SEARCH II	5	2	3	8	4	0	0	0	6	4	2	0	0	1	2	3	6	0	3	4	3	10	20.83
SEARCH I and II	6	2	3	9	6	2	1	1	6	1	0	0	1	0	3	1	7	4	0	6	1	11	22.92
SEARCH III	6	2	3	9	6	3	0	2	6	1	1	0	1	0	3	2	6	4	0	6	2	12	25.00
SEARCH IV	6	1	4	9	6	2	1	1	6	4	0	0	1	0	3	2	6	4	0	6	2	12	25.00
Faculty Impact	5	2	3	8	2	0	0	0	6	3	1	0	0	1	2	2	7	0	3	4	2	9	18.75
Project Tracking	4	2	4	8	5	1	0	1	7	3	1	0	0	2	1	3	6	0	6	2	3	11	22.92
Organisational Impact 1	3	1	3	8	1	0	0	0	5	2	1	0	0	1	1	2	8	0	3	2	2	7	14.58
Managers' Survey	1	0	3	8	2	2	0	1	0	1	0	0	0	1	0	1	10	0	3	0	1	4	8.33
Collaborative Network Evaluation	1	0	3	7	3	1	1	1	0	1	0	0	0	1	0	2	9	0	3	0	2	5	10.42

Accuracy data

The twelve items in this standard relates to technical adequacy of the report in relation to the programme under evaluation, including a determination of the merit and worth of the programme. Cumulative results are presented in Table 30.

The first item relates to programme documentation. The evaluations rated well on four of the items within this category, including the collection and description of data related to the programme, comments on how it functioned, and whether it provided a report that documented the programme operations. The evaluations rated poorly on the remainder of the programme documentation categories.

In relation to item two, which relates to context analysis, the evaluations rated poorly on all items except the use of multiple sources to describe the programme's context. There were limited references to other items that related to the overall context in which the programme was functioning, or to perception of the programme in the broader context of the health system or the stakeholders.

Item three relates to the description of purposes and procedures. The evaluations scored highly on three factors in this category: establishment of the purposes of the evaluation at the outset, recording of evaluation procedures and the use of independent external evaluators to monitor procedures.

All reports scored well on the fourth item which relates to the use of reliable and defensible information sources. There was only one category on which they scored poorly and that was the documentation of any bias in obtaining the data.

Item five relates to the validity of the information. The evaluations scored well overall in the areas of focus on the key evaluation questions, and documentation of the data collection process. However, they rated poorly in the remaining categories including training of data collectors, methods for scoring and analysing data, and comprehensiveness and categorising of the data.

The sixth and seventh items address the reliability and systematic management of the data. Item six covers the training of data collectors, use of validated measuring devices, piloting testing of methods and estimating the effects of unreliability of the data. Item seven includes establishment of protocols, use of multiple evaluators, data verification and data access. Overall the evaluations scored poorly in all of these categories. Item

eight refers to the analysis of quantitative data. There was a very limited use of quantitative data in the evaluations that were examined, and as expected they scored poorly on this item.

The ninth item refers to the analysis of qualitative data. The evaluations that reported the use of qualitative methods(8/10) scored well on the first six categories of this item. They scored poorly in the areas of assessing reliability and validity, classification of the information, establishing meaningfulness for conclusions and reporting the limitations of the methods.

The tenth item dealt with the justification for the conclusions of the evaluations. The scores in this area were disappointingly low. The only category that scored well across the reports was the accurate reflection on the procedures. There was very limited (almost non-existent) exploration of alternative conclusions and also very limited justification for the conclusions drawn. These issues relate also to item eleven which examines impartiality of reporting but also includes categories related to plausible alternative conclusions and control of bias.

Overall only three reports included an item that received an excellent rating. For the remainder, the scores were predominantly poor. The accuracy strength rating ranged from 4 to 12 with the majority (6) scoring over 10. However, this resulted in consistently low accuracy scores that ranged between 8% and 25%.

In summary these results are very disappointing. The possible reasons for this are discussed in the main body of this thesis.